Recurrent Word-Combinations in Spoken Learner English

A study of corpus data from Swedish and Norwegian advanced learners

Hege Larsson Aas

A Thesis Presented to

the Department of Literature, Area Studies and European Languages

UNIVERSITY OF OSLO

In Partial Fulfilment of the Requirements for the MA Degree

November 2011

2011
Recurrent Word-Combinations in Spoken Learner English: A study of corpus data from Swedish and Norwegian advanced learners
Hege Larsson Aas
http://www.duo.uio.no
Trykk: Reprosentralen, Universitetet i Oslo

Abstract

This project examines the inventory of recurrent word-combinations and formulaic language in corpora of native and non-native English speech, inspired by the 'corpus-driven recurrent word-combinations'-approach presented in Altenberg (1998) and De Cock (2004). The analysis draws on usage-based theories which considers recurring patterns of language to be reflective of fundamental properties of language competence. It further considers how we may best identify and describe recurring language patterns and their functions in naturally occurring speech, with a particular focus on learner language.

The primary source of material for this study is one native speaker corpus, the Louvain Corpus of Native English Conversation (LOCNEC), and two subcorpora of the non-native speaker corpus LINDSEI (the Louvain International Database of Spoken English Interlanguage), which contain speech produced by Swedish and Norwegian advanced learners of English.

The study shows some of the strengths and weaknesses of employing a hypothesis-finding, corpus-driven approach to the identification and description of formulaic language. It confirms the pervasiveness of recurrent language in both native- and non-native speech, and presents quantitative results showing how particular word-combinations are under- and overrepresented in the learner material as compared to the native speaker corpus. The more qualitatively grounded discussion draws on concepts derived from cognitive linguistics in explaining how recurrent patterns of words occur, function and change, and thus aims to position quantitative findings from corpus linguistics within this theoretical framework.

Keywords: recurrent word-combinations, advanced learner English, formulaic sequences, corpus linguistics, quantitative analysis, corpus-driven analysis, usage-based linguistics, spoken language corpora, learner corpus research, discourse markers, contrastive interlanguage analysis

Acknowledgements

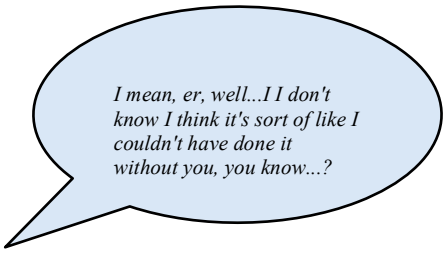
This study was conducted at the Department of Literature, Area Studies and European Languages (ILOS), Faculty of Humanities, University of Oslo.

I would like to give my thanks to a number of people who have helped me in the process of writing this thesis. First and foremost I would like to thank my supervisor Hilde Hasselgård, for her advice, encouragement and patience, for sparking my interest in learner corpus research in the first place, and for inviting me to work at the 32nd ICAME conference in Oslo this June. Attending lectures, having coffee and engaging in conversations with a large proportion of my reference list was a great source of inspiration.

I am very grateful to Sylvie De Cock at Université catholique de Louvain for granting me access to the LOCNEC corpus, to Viktoria Börjesson and Karin Aijmer at the University of Gothenburg for allowing me to study the Swedish component of LINDSEI before it was published, and to Susan Nacey and her team at Hedmark University College for giving me permission to use some of the completed interviews from the forthcoming Norwegian component of LINDSEI, and for inviting me to Hamar to take part in the process of transcribing their recorded interviews.

And to my wonderful parents, to verdens beste Josh, to my fabulous flatmates and friends, who have all believed in me (and put up with me) through this past year: *Tusen, tusen takk!*

- Hege



List of Tables and Figures

Tables

Figures

Abbreviations Used

- **NS:** Native speaker
- **NNS:** Non-native speaker
- **NL:** Native language
- **NNL:** Non-native language
- **CIA:** Contrastive Interlanguage Analysis
- **SLA:** Second Language Acquisition

***Corpora**:*

- **LOCNEC:** The Louvain Corpus of Native English Conversation
- **LINDSEI:** The Louvain International Database of Spoken English Interlanguage
    - **LINDSEI-SW,-FR,-GER,-NO:** Abbreviations for the subcorpora consisting of speech from Swedish, French, German and Norwegian native speakers
- **ICLE:** The International Corpus of Learner English
- **VESPA:** Varieties of English for Specific Purposes Database
- **BNC:** The British National Corpus
- **COCA:** Corpus of Contemporary American English

Table of Contents

1 Introduction

> "Like blood in systemic circulation, it flows through heart and periphery, nourishing all" (Ellis 2008: 9).

This study is rooted in and inspired by a set of different directions within linguistics, and attempts to reconcile these directions in the exploration of one aspect of language in use: its tendency to form and reproduce patterns. In English linguistics this aspect is most often discussed under the heading 'phraseology', which is a term encompassing work within a number of methodological and theoretical frameworks, joined together by an inclination to focus on aspects of language which have traditionally been deemed peripheral. Phraseological research is predominantly based on naturally occurring language data, or "continuous contextualized discourse" (Granger 2009: 16), often collected and compiled in a computerized database - a corpus. Gries (2009) defines a corpus as "a machine-readable collection of (spoken or written) texts that were produced in a natural communicative setting, and the collection of these texts is compiled with the intention (i) to be representative and balanced with respect to a particular linguistic variety or register or genre and (ii) to be analyzed linguistically" (Gries 2009: 411). It can be argued that a study based on such data is particularly well suited for the description and explanation of phraseological topics, since these are topics which are closely related to language in use, rather than abstract properties of language which cannot be observed through authentic speech, writing, or other means of communicative output. In *learner* corpus research, or the analysis of non-native language data, several components are brought together to form a "diversified view" (Granger 2009: 16), as illustrated in figure 1.1.:

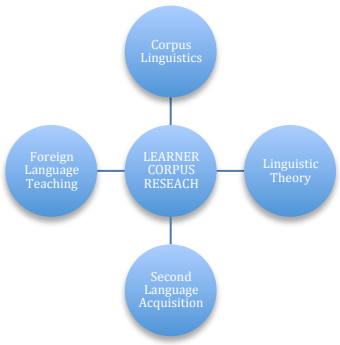


Figure 1.1 *Core components of learner corpus research (cf. Granger 2009:15)*

Since “learner corpus research lies at the crossroads between four major disciplines: corpus linguistics, linguistic theory, second language acquisition and foreign language teaching” (Granger 2009: 13), it seems necessary to consider these disciplines in relation to any analysis of learner corpus data.

This study aims to shed light on the inventory of recurrent word-combinations in native and non-native English speech, by means of a contrastive approach employing corpus data as its main source of material. It seems as if the co-occurrence and recurrence of words also lie at a crossroads, between different theoretical assumptions and methodological preferences. It is thus a further aim of this study to discuss some of these approaches, and assess their adequacy for explaining the occurrence of recurrent word-combinations, particularly those which are prevalent in learner English speech. According to Mukherjee (2009) there is a general need for “much more research into the grammar of conversation in advanced learners’ speech” (Mukherjee 2009: 226), and the Louvain International Database of Spoken English (LINDSEI) (Gilquin et al. 2010) seems to provide a suitable basis for such research, in joint action with the native-speaker corpus LOCNEC (Louvain Corpus of Native English Conversation) (ibid.).

Chapter 2 of this thesis discusses aspects of authentic communication, and argues that a study of language in general, and recurrent word-combinations in particular, is best performed on the basis of such naturally occurring language data. This chapter further accounts for “the notion that native-like proficiency in a language depends crucially on a stock of prefabricated units” (Cowie 1998: 1-2), and attempts to define the object of study based on previous research and theoretical approaches. Chapter 3 on material and method discusses whether the data in the LOCNEC and LINDSEI corpora is authentic and representative of learner and native language, and outlines the ‘corpus-driven recurrent word-combinations’-method used to extract and compare recurrent word-combinations in the two corpora. Chapter 4 presents the quantitative analysis of the material and preliminary discussions, while chapter 5 delves further into the nature of recurrent word-combinations in relation to the quantitative and qualitative findings of the preceding chapter.

2 Recurrent Word-Combinations in Spoken Learner Language: Theoretical background

2.1 Research Based on Naturally Occurring Spoken Language Data

2.1.1 Introduction

Setting out to understand what can be gained from research based on a corpus of spoken language data, it is perhaps useful to start by trying to define what spoken language really is and how it relates to written language as well as to the wider concept of 'language itself'. In his article "The spoken language corpus: a foundation for grammatical theory", Michael Halliday explains why he thinks "the essential nature of language (...) is most clearly revealed in the unselfconscious activity of speaking" (Halliday 2004a: 25):

> "This is where systematic patterns are established and maintained; where new, instantial patterns are all the time being created; and where the instantial can become systemic, not (as is more typical of written language) by way of single instances that carry exceptional value, but through the quantitative effects of large numbers of unnoticed and unremembered sayings" (ibid.).

Before the introduction of spoken language corpora these 'systemic patterns' of spoken language were hard to identify, and initially it was, in Halliday's words, "the tape recorder that broke through the sound barrier (the barrier to arresting speech sound, that is) and made the enterprise of spoken language research possible" (ibid.: 11). Later, technological advances in the development of personal computers and computer software have proven very important for gathering natural language data, both spoken and written. Today, compiling a corpus of spoken language is time consuming and involves several methodological difficulties (cf. chapter 3 on material and method), but through the combination of computerized and manual work corpus linguistics has become an invaluable method for the study of spoken language, a scenario which must have been difficult to imagine only a few decades ago.

However, the biggest obstacle to overcome in placing patterns of language use on the linguistic agenda was perhaps not technical or methodological, but theoretical and ideological, and I will in this section discuss some of the motivations behind performing research based on authentic language data in general, and spoken

language data in particular. What theoretical assumptions is a study of language use typically based on, what conclusions can justifiably be drawn from such 'usage-based' corpus research, and what are, in this respect, the advantages of studying spoken language as opposed to written language, if any? These are, I believe, important considerations to keep in mind when performing a study such as the present one, if its results are to be considered meaningful. The discussion mainly concerns native language usage, but is equally applicable for the discussion of the behaviour of learner language and the learner's journey towards "nativelike selection and nativelike fluency" (Pawley and Syder 1983). In light of the following discussion, I will also argue that the study of authentic spoken language data is essential for a better understanding of formulaic language and recurrent word-combinations in both native and learner language.

2.1.2 Towards a Corpus-Friendly Theory

Noam Chomsky's early publications on syntactic theory, *Syntactic Structures* (1957) and *Aspects of the Theory of Syntax* (1965), revolutionary within the field of linguistics, have inspired and influenced linguistic research for decades. Chomsky's idea of a mental "system of rules determining the interpretation of its infinitely many sentences" (Chomsky 1965: v), the foundation for generative grammar, led to a new and narrowed definition of the true object of study in linguistics, namely that of a set of complex mental structures making it possible for us to understand, and produce, language, i.e. "the speaker-hearer's *knowledge* of his language" (ibid.: 4, my italics). 'Performance', or language use, with its "numerous false starts, deviations from rules, changes of plan in mid-course, and so on" (ibid.), was seen as the product of more factors than language competence alone, and thus "in actual fact (...) obviously could not directly reflect *competence*" (ibid., my italics), which ought to be the sole object of language studies. Furthermore, Chomsky's notion of competence stipulates a mental lexicon which contains no redundancy and only the non-predictable features of language, i.e. only the lexical items needed in order to construct, or generate, sentences according to the rules of grammar (Jackendoff 2002: 153). This combination of mental lexicon and rules thus makes us capable of both constructing and comprehending an indefinite number of novel utterances, and makes for a powerful, creative system.

Chomsky's view was later criticised from a number of perspectives, and several linguists today believe that a strict distinction between language use and the more abstract language knowledge or competence has concealed features vital to their understanding of what language competence must comprise (if indeed there is such a separate component for language competence in our brains). The identification of highly recurrent patterns of combined words in language use challenges the generative notion that lexical items are combined freely with no restrictions other than the ones posed by grammatical rules. In addition, it is not easy to adequately explain these observed patterns in terms of vague definitions of a separate 'performance':

> "Whereas it was previously possible to imagine that words combined fairly freely, their restrictions attributable to context and pragmatics, and to easily definable social signalling, it is now clear that, once you actually map out the patterns of distribution for words, no such piecemeal and superimposed explanation is possible. Words belong with other words not as an afterthought but at the most fundamental level" (Wray 2002: 13).

It thus seems that corpus studies of language use which reveal frequency patterns of collocation are not compatible with the generative theoretical framework, and need to seek a theoretical foundation elsewhere. Most generativists believe that "important insights and generalizations about linguistic structure may be missed if vague semantic clues are followed too closely" (Chomsky 1957: 101), and the importance of general patterns has continued to be a prevailing notion in e.g. the teaching of grammar in schools (cf. Sinclair 1999a). Evidence from authentic language data, however, show a linguistic reality far from the structurally creative scene proposed by traditional generativism:

> "Native speakers do *not* exercise the creative potential of syntactic rules to anything like their full extent, and (…) indeed, if they did so they would not be accepted as exhibiting nativelike control of the language. The fact is that only a small proportion of the total set of grammatical sentences are nativelike in form – in the sense of being readily acceptable to native informants as ordinary, natural forms of expression, in contrast to expressions that are grammatical but are judged to be 'unidiomatic', 'odd' or 'foreignisms'" (Pawley and Syder 1983: 193).

Acknowledging that "grammar is more specific and restricted than the simple rewrite rules of mainstream generative models claim" (Wray 2009: 81) there is thus a call for a different approach to language processing and acquisition, which is capable of explaining performance observations within a more comprehensive framework.

2.1.2.1 'Performance', 'parole' and 'usage'

The question of where the focus of linguistics should be was addressed a long time before Chomsky's *Syntactic Structures* was published. In *Language*, Leonard Bloomfield establishes initially that "the most difficult step in the study of language is the first step" (Bloomfield 1933/1967: 21), and in Switzerland two decades earlier, Ferdinand de Saussure's book *Course in General Linguistics* asked: "What is it that linguistics sets out to analyse? What is the actual object of study in its entirety?" (de Saussure 1915/1983: 8). According to Saussure, we ought to understand language introspectively through a theory of underlying mental and social structures, as studying empirical language data would ultimately result in "a muddle of disparate, unconnected things" (ibid.: 9). Following this, Chomsky's ideas some fifty years later seem to be compatible with Saussure's, but Saussure's concept of the 'language itself' or, in his words, 'langue', differs in some respects from Chomsky's concept of 'competence', and where it differs we can find some stimulus for a theory of language based on language use:

> "It [langue] is a fund accumulated by the members of the community through the practice of speech, a grammatical system existing potentially in every brain, or more exactly in the brains of a group of individuals; for the language is never complete in any single individual, but exists perfectly only in the collectivity" (ibid.: 13).

There are similarities between these ideas and Halliday's belief in systematic patterns that are constantly being created through 'the quantitative effects of large numbers of unnoticed and unremembered sayings'. When assuming that language use both reflects and affects our mental ability to understand and produce language, as well as our language output, it also seems reasonable to consider performance as a useful or even crucial object of study, even if one is aiming to gain insight into theories about linguistic knowledge or 'competence'. What is often referred to as 'mentalist' theories, such as Chomsky's generativism, may, some linguists claim (cf. Dyvik 1995), run the risk of *creating* its object of study, i.e. abstract mental structures, and thus becoming a theory of these abstract relations rather than a theory which serves to explain language comprehension and production to its full extent: "if we confine our attention to language instead of the actual speech acts that embody a use of language, there is, quite literally, nothing happening" (Dretske 1974: 24; cited in Dyvik 1995: 30). Similarly, John Sinclair argues that theories based on intuition are ultimately

theories about the nature of intuition, not about the nature of language: "human intuition about language is highly specific, and not at all a good guide to what actually happens when the same people actually use the language" (Sinclair 1991: 4).

In his work *Sociolinguistic Patterns* from 1972, William Labov drew further attention to linguistic theories which aim to explain *all* aspects of natural language use data, rather than excluding those aspects from consideration which do not fit neatly into a rule-based generative framework. Intuitive methods are also valuable for the study of language which is defined this way, Labov claims, and there are advantages arising from identifying general structures, but on their own these methods and the underlying theory prove to be inadequate:

> "We cannot afford any backward steps: anyone who would go further in the study of language must certainly be able to work at this level of abstraction. At the same time, it is difficult to avoid the common-sense conclusion that the object of linguistics must ultimately be the instrument of communication used by the speech community; and if we are not talking about *that* language, there is something trivial in our proceeding" (Labov 1972: 187).

Sociolinguistic Patterns thus presents methods for establishing abstract relations between performance data, and claims that "it is reasonable to believe that they are more than constructions of the analyst – that they are properties of language itself" (ibid.: 259). In this way, performance is being included in a wider concept of competence, one that can be explored through both introspective methods and the behavioural study of natural language use data.

2.1.2.2 'Language in its entirety': Cognitive linguistics and corpus linguistics

The assumption that studies of language use are crucial for the development of theories about 'language itself' is fundamental within several branches of linguistics, which, consequently, may be labelled 'usage-based'. In this context, the term 'usage-based' covers approaches and theories which are not only methodologically based on language use data, but which also presuppose that language is in some way shaped by usage – a view which contradicts the less flexible and autonomous rules of generativism. A usage-based approach which has gained ground in recent decades is the cognitive linguistics approach, which sees our general cognitive capacities and our capacity for language as dynamic entities, shaped by "cognition, consciousness, experience, embodiment, brain, self, and human interaction, society, culture and

history" (Ellis and Robinson 2009: 3). When encouraged to "think of language as ever being affected by language use and the impact that experience has on the cognitive system" (Bybee 2010: 4), analysing language use data is recognized in a broader perspective in cognitive theory, where findings which cannot conform to generalized rules are no longer disregarded as trivial and insignificant. The cognitive linguistic framework thus makes it theoretically justifiable to explain phenomena such as semantic and pragmatic change, as well as language patterning and recurrent word-combinations, on the basis of e.g. increased frequency of use (cf. Bybee 2010). In this perspective, form, meaning and context is considered to be part of out mental representation of language, and there is thus "no separation of 'surface' and 'underlying' levels, since "it is difficult, when actual data is considered, to draw a clear boundary between lexis and structure" (Barlow 1996: 22). As pointed out by e.g. Gries (2009: 407): "many of the assumptions underlying cognitive-linguistic work—in particular the relevance assigned to frequency of patterns in learners' input (…) are often completely analogous to working assumptions in Corpus Linguistics" (ibid.), and the past years have seen calls for combining descriptive corpus studies and theories of the mind: "to assume as the main theoretical framework within which to explain and embed our analyses a psycholinguistically informed, (cognitively-inspired) usage-based linguistics" (Gries 2010a: 338).

A cognitive foundation may also be found in Wallace Chafe's (1992) idea of what a "corpus linguist" should be: "(...) it is a linguist who tries to understand language, and behind language the mind, by carefully observing extensive natural samples of it and then, with insight and imagination, constructing plausible understandings that encompass and explain those observations" (Chafe 1992: 90). Similarly, Leech (2000) emphasises the necessary connection between theory and data postulated by most researchers working with corpora: "corpus linguists assume that relevant theories or hypotheses must be capable of confirmation or disconfirmation through empirical observation of language in use" (Leech 200: 685). The term 'corpus linguist' followed the technological development of computers and corpora to describe linguists who made use of these new possibilities, which makes it reasonable to refer to 'corpus linguistics' as a methodology rather than a theory (Meyer 2002: xi). Chafe's definition, however, underlines the fact that "there is in principle no conflict between

being a corpus linguist and being a theoretical linguist" (Aarts 2000: 7), and also opens up for a combination of methods irrespective of initial ideologies, and a "necessary culture of collaboration" (Jackendoff 2002: 429) between the fields of linguistics in a common quest for understanding the nature of language, which in turn is important for the applying of linguistic knowledge into fields like second language acquisition research. Today there are still discussions on the value of empirical language use data, both as an object of description and as a means of expanding our knowledge of language in the mind. "You want an answer to a non-trivial question, you've got to go beyond looking at data" (interview with Noam Chomsky; Aarts 2000: 6), Chomsky objects, and points to one of the dangers of corpus linguistics – the so-called 'number-crunching' or presentation of frequency data with no prior theory or subsequent analysis. However, this objection ignores the fact that corpus-based studies may very well avoid this pitfall, and rather use the collected data to prove, disprove or create theoretically valid assumptions. As emphasized by e.g. Barlow: "the problem of extrapolating from the corpus to 'language' is always present" (Barlow 1996: 2), and corpus studies and results run the risk of illustrating Saussure's 'muddle of disparate, unconnected things'. However, a common feature of theories based on language use, such as Halliday's systemic functional linguistics, is their aim to be "comprehensive" and "(…) concerned with language in its entirety, so that whatever is said about one aspect is to be understood always with reference to the total picture" (Halliday 2004b: 19). This focus on 'language in its entirety' seems to be one shared by most linguists working with corpora, and which thus directly or indirectly influences analyses, conclusions and applications.

2.1.2.3 Corpora and applied linguistics

Excluding language performance from consideration also affects fields of applied linguistics, such as language pedagogy and studies of second language acquisition, in an unfavourable way, and Sinclair (1999a) claims in his article on frequent words and their patterns of usage, *A Way With Common Words*, that "language teachers have become apologists for an inadequate model of the language" (Sinclair 1999a: 159), largely owing to the principles of generative grammar. This "grossly oversimplified" (Sinclair 1999b: 2) model, postulating a strict separation of syntax and semantics, thus makes it difficult to create meaningful descriptions and explanations when having to

do with authentic language use. Similarly, Mukherjee (2006), writing on the use of corpora in language teaching, places emphasis on findings which indicate that "the scope of virtually all grammatical rules is limited and that there is a remainder of instances which deviate from the rules" (Mukherjee 2006: 11), and believes the generative framework to be unable to account for these instances. Sinclair (1991), on the basis of corpus evidence, introduces an alternative model to the 'slot-and-filler' model promoted by generative grammar, a model which softens the syntax-lexis distinction and thus is suitable for the description of features such as language patterning and phrasal semantics. From Sinclair's point of view, corpora make it difficult to defend and uphold models which cannot adequately explain authentic language: "It is now rather risky to make faulty statements about a language, since access to corpus evidence is getting easier every day" (Sinclair 1999b: 14). Intuitive ideas and 'faulty statements' are thus put to the test through the increasingly large data on linguistic behaviour that we have at our disposal, which makes for descriptions of an increasingly objective (and non-prescriptive) nature:

> "when traditional descriptions have included guidance on language use, it is usually based on the author's perception of appropriateness. In contrast, corpus-based analysis allows us to discover what typical speakers and writers actually do with the grammatical resources of English" (Biber and Reppen 1998: 145).

Leech (2000) argues that evidence from corpora, and the subsequent development of "*performance grammars*" (Leech 2000: 686), is particularly valuable in the context of language learning: "it is difficult to suppose that we could learn to use the grammar of a language effectively without being attuned to the conditions and constraints determining its use" (ibid.). Thus, corpus data becomes important not only as means of investigating the processes of language use and acquisition, but also as descriptive data which may provide helpful tools for learners. The authentic setting of corpus data also provides information on situational content, and it is thus possible to develop a grammar which "takes account not only of the self-contained grammar system of a language, but how external considerations determine choices from the system, and how the system relates to other aspects of linguistic communication" (Leech 2000: 687). It thus seems that corpus linguistics, usage-based theories and second language acquisition research and the application of these theories, ought to go hand in hand in the future. This fact is also stressed by Stig Johansson (2009):

> "It seems to me that the usage-based model and the relevance of corpora deserve to be recognised in works on second-language acquisition. I have looked at some recent works to see if there was anything to find on the role of corpora. I found preciously little. And yet there is a discussion of notions which are central in the use of corpora for teaching and learning, such as attention and awareness, input, hypothesis formation. Here is a task for the future for those who believe in the validity of the usage-based model and the corpus-based approach" (Johansson 2009: 39).

2.1.3 Properties of Spoken Language

2.1.3.1 Spoken and written grammars

Following a move from a traditional generative to usage-based approaches, along with the development of computer tools, several corpora of written language have been compiled and many studies have been performed on the basis of this material to investigate our written language behaviour. However, as mentioned above, greater technical and practical difficulties have long prevented systematic corpus studies of *spoken* language. This has had implications for accounts of the character of spoken language, and it may be said that yet again, a significant component of language was disregarded. Because of the lack of sufficiently large-scale empirical studies to counter this approach, smaller samples of spoken language have often been subject to analyses according to the grammatical standards of written language:

> "(…) it may be argued (…) that the models of *grammar* which underpin most of the laudable attempts at representing and activating the use of the spoken language are still rooted in descriptions of the grammar of written English and have failed to take on board some interesting features of the grammar of informal, interactive talk" (Carter & McCarthy 1995: 141).

This bias toward written language, Carter and McCarthy believe, might result in a view amongst descriptive grammarians and laypeople alike, that features displayed in samples of spoken language are 'wrong' in the sense of not conforming to the rules of grammars based on written language:

> "(…) written-based grammars exclude features that occur widely in the conversation of native speakers of English, across speakers of different ages, sexes, dialect groups, and social classes, and with a frequency and distribution that simply cannot be dismissed as aberration. If our speakers are 'wrong', then most of us spend a lot of our time being 'wrong'" (Carter and McCarthy 1995: 142).

Naturally, most usage-based studies today are cautious about labelling lexical or syntactic choices as 'right' or 'wrong' in a prescriptive fashion, but a lack of

descriptions of spoken language is still unfortunate in that it prevents us from uncovering the systematic features within it: "precisely because there are patterns which don't occur in writing, we need a corpus of spoken language to reveal them" (Halliday 2004a: 19). Gilquin and De Cock (2011) also point to the fact that studies of spoken language often reveals other features of spoken language apart from lexis and grammar, such as "the presence and the functions of fillers, pauses and other related phenomena" (Gilquin and De Cock 2011: 147). Similarly, Aijmer (2011) stresses how research based on authentic speech "goes hand in hand with discovering that repeats, pauses, false starts, pragmatic markers are not 'errors' but are part and parcel of 'conversational grammar'" (Aijmer 2011: 232). According to Chafe, "speaking is natural to the human organism in ways that writing can never be" (Chafe 1992: 88), and even though electronic media such as chats and blogs, "hybrid varieties of language" (Wray 2009: 52), are increasingly moving closer to the spoken mode, there seems to be little doubt that the situational context of spoken language still holds certain characteristics which cannot to the same extent be found in the context of written language, such as very low levels of self-monitoring. Before spoken language corpora "written text had to serve as the window, not just into written language but into language" (Chafe 1992: 13), and the inclusion of spoken language data in research seems to be all-important for our understanding of language in general.

Some recent corpus based reference grammars, such as the Longman Grammar of Spoken and Written English (LGSWE) (Biber et al. 1999), include sections on selected registers of spoken language, as well as possible explanations for the differences found between the language of various spoken and written registers. Several researchers, e.g. Halliday (2004b), are opposing the idea of separate grammars for spoken and written language, believing it creates an artificial absolute difference between the two modes:

> "This approach [writing a separate grammar for spontaneous speech] has the merit that it can highlight special features of spoken language and show that it is systematic and highly organized; but it tends to exaggerate the difference between speech and writing, and to obscure the fact that they are varieties within a unitary system. Spoken and written English are both forms of English – otherwise you could not have all the mixed and intermediate forms that are evolving in electronic text" (Halliday 2004b: 34).

In LGSWE, Biber et al. (1999) seem to agree with this "underlying *sameness* of spoken and written grammar" (Leech 2000: 687), and place emphasis on the multitude of factors that may have an impact on our linguistic choices, "such as the reason for the communication, the context, the people with whom we are communicating, and whether we are speaking or writing" (Biber et al. 1999: 5). Similarly, Wray (2009) believes that "the fundamental differences between spoken and written language are secondary, rather than primary" (Wray 2009: 57), and that it is the needs and aims of the communicators which ultimately decides linguistic output: "The medium of expression facilitates, rather than determines, the differences between text types" (ibid.: 57-58). The mode of communication is thus considered to be one of several such factors, more or less influential depending on the other factors composing a register. Following this, the appearance of features in a spoken or written register is determined on a probabilistic scale, creating a "unified model" (Leech 2000: 692) of speech and writing. Leech (2000) concludes that "there is comfort for the learner (…) in the conclusion that, for practical purposes, [native speakers] and [non-native speakers] have one grammatical competence to acquire, not two" (ibid.: 714). However, perhaps even more challenging for the learner are register differences, and the command of language features which signal "some shared, underlying communicative functions associated with the situational contexts of the texts." (Biber 1993: 335). This difficulty is often pointed out in studies of features of learner language, particularly in writing, where certain features are often said to create an impression of a "speech-like nature" (Granger and Rayson 1998: 129). The probabilistic scale of language features is thus partly determined by our knowledge about certain registers, and by what characterises them in terms of their situational demands, limitations and possibilities.

2.1.3.2 'Unself-monitored spontaneous speech' as a new window into language, and characteristics of spoken conversation

Performing research based on spoken language data is thus important for the creation of a more precise grammar, taking into account a broader range of our linguistic repertoire, and for further rejecting the view that spoken language per definition is incoherent, unsystematic and prone to errors. However, even though the strict boundaries between spoken and written language may be erased, it seems that certain

spoken registers are particularly valuable for the study of e.g. language learning and formulaic language, as they may provide a particularly close insight into processes of language production and comprehension. Returning to Halliday's claim that "the essential nature of language (...) is most clearly revealed in the unselfconscious activity of speaking", we find motivation for studying these spoken language registers:

> "(...) it is in the most unself-monitored spontaneous speech that people explore and expand their meaning potential. It is here that we reach the semantic frontiers of language and get a sense of the directions in which its grammar is moving" (Halliday 2004b: 34).

Chafe (1992) claims that "(…) the collection and analysis of conversational corpora is absolutely essential to a fuller understanding of language and the mind" (Chafe 1992: 89), which can be said to be due to certain characteristics of the register conversational corpora represent. The register of spoken conversation, the "quintessential spoken variety" Leech (2000: 690), is considered to be "frequently very different" (ibid.) from other registers, due to the situational demands of the register. Table 2.1 presents an outline of external determinants of spoken conversation as proposed by Biber et al. in the LGSWE (1999: 1041-1052):

i.	Conversation takes place in the spoken medium
ii.	Conversation takes place in shared context
iii.	Conversation avoids elaboration or specification of meaning
iv.	Conversation is interactive
v.	Conversation is expressive of politeness, emotion and attitude
vi.	Conversation takes place in real time
vii.	Conversation has a restricted and repetitive repertoire
viii.	Conversation employs a vernacular range of expression

***Table 2.1**: External determinants of conversation (cf. Biber et al. 1999: 1041-1052)*

Biber et al. (1999) employ these determinants as explanations for differences found between a corpus of spoken conversation and corpora of written registers, and some of the determinants (e.g. vi.) are particularly relevant as features which explains why spoken conversation data can give us insight into language processing in ways which more monitored written registers cannot: "unlike written registers, conversation

suffers from the pressures of real-time processing, bringing overload on the short-term (working) memory" (Leech 2000: 698). Leech (2000) claims that "conversational grammar is adapted to the needs of real-time processing" (Leech 2000: 698), and that "on-line pressures encourage reliance on a limited repertoire of items readily retrievable from memory" (ibid.: 701). Similarly, Altenberg and Eeg-Olofsson (1990) brings attention to how "unlike writers and speakers reading from a manuscript, spontaneous speakers must plan, encode and execute their utterances in real time, i.e. extremely fast and almost simultaneously" (Altenberg and Eeg-Olofsson 1990: 1). The connection between e.g. on-line pressures and linguistic output in spoken conversation is illustrated in figure 2.1 below, which picks up on most of the determinants of table 2.1:

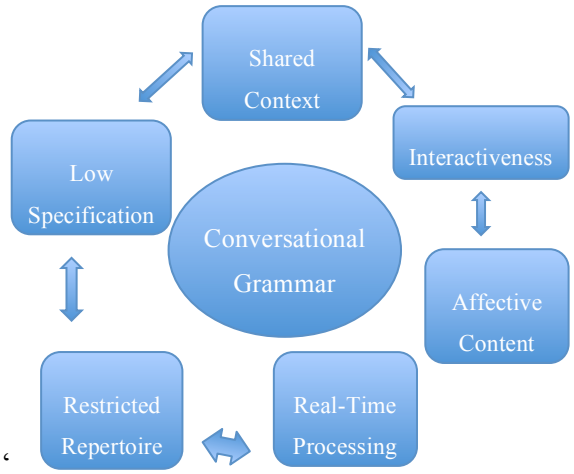


Figure 2.1: *The interrelated functions associated with conversational grammar (cf. Leech 2000:701)*

Figure 2.1 sees a 'restricted repertoire' as a direct consequence of the real-time processing of conversational grammar. This entails, a greater reliance on few, but recurrent words and word-combinations in spoken conversation as opposed to written registers, in agreement with the findings of e.g. Biber et al.: "Time pressure makes it more difficult for speakers to exploit the full innovative power of grammar and lexicon: instead, they rely heavily on well-worn, prefabricated word-sequences, readily accessible from memory" (Biber et al.: 1049). This connection is also made by e.g. Ellis et al. (2008: 376) and Altenberg and Eeg-Olofsson (1990): "(...) spontaneous speakers frequently resort to repetition and more or less mechanical 'recycling' of

stored or previously used expressions to simplify the task of production, to render discourse more coherent, or to realize particular conversational strategies" (Altenberg and Eeg-Olofsson 1990: 1-2). These approaches thus agree with Wray's (2009) emphasis on how the needs of the speaker, and hearer, are facilitators for the occurrence of recurrent word-combinations in language, assessing what is "advantageous during linguistic processing" (Wray 2009: 69) at any given time. In general "there seems to be a link between the use of formulaic sequences and a need and desire to interact" (Wray 2002: 175), which is particularly relevant for spoken, face-to-face situations, and which must be assessed along with the constraints and possibilities of these situations, as illustrated in table 2.1 and figure 2.1. It is likely that the balance between "the achieving of successful interactional events and the saving of processing effort" (ibid.: 198), is an important determinant for the 'restricted repertoire' of spoken language, as a restricted and conventionalised linguistic output is likely to facilitate conversation both from a hearer and speaker perspective. For the purpose of this study, it is also assumed that there is a greater reliance on fewer word-combinations in learner speech, as the on-line pressures on the learner as caused by e.g. lexical retrieval difficulties should generally be greater than the processing efforts on native speakers: "more planning is needed and more monitoring, and therefore more time" (Dechert 1984: 224). Recurrent word-combinations found in interlanguage speech may thus be the result of a wish to create fluency and reduce processing load for more complex language, or language perceived to be problematic from the learner's point of view. In addition, learners may have a more restricted inventory of lexical choices in general, thus making for an even more repetitive language, dominated by certain 'islands of reliability', "points of fixation, anchoring grounds to start from and return to!" (Dechert 1984: 223).

2.1.4 Research Based on Naturally Occurring Spoken Language Data: Summary

As stated in the introduction to this chapter, it seems as if the use of a spoken corpus, particularly a corpus of spoken conversation, is ideal for the study of external speech, internal language competence and situational factors of communication; and of how these three components are linked together. The next chapter will focus on the object of study in this thesis; recurrent word-combinations and evidence of formulaicity in spoken learner language. This analysis is founded on a "commitment to a particular

view of language processing, namely, that it is the accessing of large prefabricated chunks, and not the formulation and analysis of novel strings, that predominates in normal language processing" (Wray 2002: 101). It is possible that without the study of authentic data in general and electronic corpora in particular, recurrent patterns would not have been recognized as important for our understanding of language or language learning at all. Word-frequency and language patterning are aspects of language that can be hard to pin down, define, and incorporate into theory, as will be demonstrated in the next section, but this is not a valid reason to ignore them. Today, linguistic theory and models seem to be taking into account the occurrence of recurrent patterns of language, "no doubt because corpus linguistic research cannot be ignored and it finds them ubiquitous" (Wray 2009: 87). Evidence of the complexity of language and the focus on word-combinations which has been brought to the forefront by corpus linguists and others thus hold the potential to create "a better understanding of language, be it in terms of cognition, description, acquisition or teaching" (Granger and Meunier 2008:15), even if it means moving beyond orderly categories. As summed up by Wray (2009):

> "In many different ways, we are at the boundaries - of language behaviour, of communicative potential, and of linguistic theory - and aim to see what happens when squeeze a phenomenon until (as we say formulaically) the pips squeak" (Wray 2009: 5).

2.2 Recurrent Word-Combinations and Formulaic Sequences in an Interlanguage Perspective

2.2.1 Introduction

The linguistic occurrences this thesis sets out to investigate are multifaceted, both in terms of their many and diverse definitions, and considering the many terms that are used to describe them by different researchers. As seen in section 2.1 above, evidence of language patterning found in studies of language use have been found to contradict generative models, and to be part of language itself, "not as an afterthought but at the most fundamental level" (Wray 2002: 13). This shift in focus from peripheral to fundamental has been apparent in several publications in recent years, and there is a general consensus in many linguistic circles that while knowledge about abstract rules of grammar might play a part in language production and comprehension, our ability to understand and produce longer, unanalyzed stretches of language is equally important, leaving traditional descriptions of grammar responsible for an adequate explanation of this duality: "If the native speaker knows certain linguistic forms in two ways, both as lexical units and as products of syntactic rules, then the grammarian is obliged to describe both kinds of knowledge; anything less would be incomplete" (Pawley & Syder 1983: 217). There are still numerous questions to be asked about these collocational patterns, such as how they are acquired and stored, and, as will be the primary concern of the present thesis; how they function in texts, how we can properly identify them as in some way being 'formulaic', and how they may contribute to differences found between native language data and learner language data. These are questions which bring up both externally observable features of text as well as hypotheses about features internal to our linguistic system. Issues regarding identification, definition and terminology will be discussed in section 2.2.2.

In studies of second language acquisition and interlanguage, as in studies of language in general, evidence of language patterning have been largely ignored, which is also, according to Wray: "due, in part, to a theoretical bias towards looking at lexis and grammar separately and assuming that language learning primarily entails the building up of larger units from smaller ones" (Wray 2002: 173). However, as several studies of both native and second language questioning this "theoretical bias" have

appeared in recent years, frequently making use of corpus methodology, we have started to gain increased knowledge about the patterning of language. Many researchers have come to believe that these patterns are "not only useful for efficient language use; they are essential for appropriate language use" (Schmitt & Carter 2004: 10), and so including this knowledge in our accounts of interlanguage and second language acquisition theory seems pivotal. Indeed, attempts are increasingly made to answer e.g. Gries's call for "integrating […] accounts of phraseologisms in particular and other patterns more generally into a larger theory of the linguistic system" (Gries 2008: 22), and these accounts should also have implications for our view of the linguistic system in relation to acquisition theory. Whether our linguistic systems differ in terms of native and second language processing is important for our understanding of language patterning in interlanguage data as compared to native language data, and section 2.2.3 will discuss some relevant issues related to these presumed differences, as well as properties of interlanguage and the language learner related to this study. The researchers behind the LINDSEI and LOCNEC corpora conclude that "phraseological errors are universally recognized as those that most clearly distinguish native from non-native language, even at an advanced proficiency level" [1], and several studies conducted in recent years have been initiated with this assumption in mind. As the subjects in the LINDSEI corpus are all such advanced learners of English, the overview on previous research of recurrent word-combinations in learner language in section 2.2.4 will center on studies relevant to this particular learner population.

2.2.2 'Recurrent word-combinations', 'Formulaic sequences', 'Chunks', 'Patterns', 'Units', 'Prefabs' and 'Phraseologisms': Defining the Object of Study

Many will agree with Bengt Altenberg that phraseology, or the study of phrases, is "a fuzzy part of language" (Altenberg 1998: 101), and the multitude of terminology and definitions that is found in research literature on phraseological topics cannot be said

[1] University of Louvain, *Foreign Language Learning: Phraseology and Discourse Action de recherche concertée* [URL], http://sites-test.uclouvain.be/cecl/projects/PhraseologyARC/welcome.html

to help make matters clearer. This development is unfortunate, since it makes the field appear more scattered than it actually is, and may thus cause unnecessary delay in the effort to include language patterning as a component of more comprehensive linguistic theories, and in the process of applying it in e.g. second language acquisition contexts. Similarly, difficulties may arise in the analysis of word-combinations and the replication and comparison of studies if definitions prove to be lacking, inconsistent or without theoretical foundation. In an attempt to combat the muddle of terms and definitions, Gries (2008) proposes six parameters which ought to be used as clarification of our choices of objects of study, since "it is essential that we, who are interested in something as flexible as patterns of co-occurrence, always make our choice of parameter settings maximally explicit to facilitate both the understanding and communication of our work" (Gries 2008: 10). However, these parameters need not necessarily be identified prior to the identification of patterns: "Sometimes it may be more revealing to let the data – rather than the preconceptions of any particular researcher – decide what the potentially most revealing pattern is" (ibid.: 21). This thesis will initially approach patterning of language in such a 'bottom-up' manner, and has thus adopted the term 'recurrent word-combinations' from Altenberg (1998), defined as "any continuous string of words occurring more than once in identical form" (Altenberg 1998: 101). According to Altenberg, this is a "rather non-committal approach" (ibid.), but, as mentioned by Gries, it is an approach which may reveal patterns that would otherwise be overlooked in a more tailored search. A search for particular patterns may similarly fail to notice important differences between corpora in a comparative study such as the present one. In addition, as mentioned by e.g. DeCock (2004: 227-228), Read and Nation (2004: 30), the difficulties involved in intuitively constructing a pre-established list of conventional patterns in non-native language make the n-gram approach particularly suitable for the analysis of language patterning in interlanguage data.

It is, however, my intention to identify such sequences in the LINDSEI and LOCNEC data which may be considered as potential multi-word *units*, stored and retrieved whole and used to display (a) particular, unified function(s) in the text, and thus some of the n-grams retrieved from the frequency search will eventually be discarded from consideration. This process may be problematic in theoretical terms and undermine

the objective stance of the initial search, as such cutting down on research material is often based on intuitive and arbitrary decisions (cf. Wray 2002: 26-27). However, a clear definition according to e.g. Gries' parameters, as well as clear justifications as to what is left out of the analysis, may help retain objectivity and comparability. I will return to what De Cock (2004) terms "the corpus driven 'recurrent word combination' method" (De Cock 2004: 227) in further detail in chapter 3 on method, and devote the following sections to a presentation of Gries' definition of word-combinations according to six parameters, as well as Alison Wray's (2002) broader definition of 'formulaic sequences', which takes a more internal approach to the matter. Other understandings of language patterning will also be discussed, including accounts of the phenomena from the perspective of usage based cognitive linguistics. Section 2.2.2.3 will discuss the approach I have chosen for the present thesis in terms of definition and terminology.

2.2.2.1 Word-combinations defined according to six parameters

Following his own parameters, presented in table 2.2 below, Gries defines his notion of a 'phraseologism':

> "a phraseologism is defined as the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of occurrence is larger than expected on the basis of chance" (Gries 2008: 6).

i.	the *nature* of the elements involved in a phraseologism;
ii.	the *number* of elements involved in a phraseologism;
iii.	the *number of times* an expression must be observed before it counts as a phraseologism;
iv.	the permissible *distance* between the elements involved in a phraseologism;
v.	the degree of *lexical and syntactic flexibility* of the elements involved;
vi.	the role that *semantic unity* and *semantic non-compositionality/non-predictability* play in the definition.

***Table 2.2**: Six parameters for defining a phraseologism (cf. Gries 2008: 4)*

Some of the parameters are more controversial and open for discussion than others, although there is commonly broader agreement about the alternative approaches among researchers working within similar fields and employing similar methods, such as the corpus methodology. Although specification of criteria is the most important

issue, in that it allows for closer scrutiny of studies of language patterns as well as grounds for replication and comparative studies, a definition may also include assumptions about the nature of word-combinations which cannot easily and objectively be determined according to Gries' parameters. I will argue that such assumptions have implications for the overall validity and importance of studies of word-combinations, and that a definition should thus make reference to them, whether it be implicit or explicit.

2.2.2.2 Frequency, unity and prefabrication of word-combinations

Gries' definition of a phraseologism comprises the most important initial criteria of the present analysis, namely frequency of occurrence, as this is an inherent concept in the corpus driven recurrent word-combinations-method. It is commonly held in corpus linguistics that a highly frequent word-combination must in some way be conventionalised by the speech community, and possess certain qualities which makes it a central subject of analysis. However, as will be seen in the extraction of n-grams from LINDSEI and LOCNEC, not all frequent patterns can, based on intuition, be said to belong to conventionalised speech, or to be in any other respect considered unified. To exclude these frequent but not unified word-combinations, Gries demands semantic unity in his definition, and further adds: "it is probably fair to say that there is little work which has defined phraseologisms solely on the basis of some quantitative criterion based on their frequency of occurrence" (ibid.: 5). Semantic unity, "to have a sense just like a single morpheme or a word" (ibid.: 6) is in Gries' parameters separated from semantic non-compositionality, where the meaning of the phraseologism cannot be inferred on the basis of grammatical and semantic analysis of its separate components. I will get back to the definition employed in my classification of the n-gram results below, where compositionality is considered to be an important factor, in addition to a certain degree of semantic/pragmatic unity.

On the other hand of the frequency scale, Gries' definition does not include all instances of what may be viewed as 'archetypal' language patterns such as idioms and proverbs, as these word-combinations, though semantically unified, often display relatively *low* frequencies in authentic speech and writing (Biber et al. 1999: 989; Moon 1998). These are patterns which are intuitively recognized as somehow

belonging together, since they are likely to be memorized as wholes, as well as typically being semantically unified and non-compositional. In cognitive linguistics, it is commonly presumed that the frequency of a 'unit' (cf. Gries 2008: 13) determines its accessibility in production and comprehension since it leads to pattern entrenchment, but, as argued by e.g. Bybee (2009), units do not have to be particularly frequent in order to be sufficiently memorized, as in the case of idioms. This fact is also emphasized by Wray (2002), who does not include frequency information in her broader definition of 'formulaic sequences':

> "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray 2002: 9).

Wray (2002) further concludes, on the basis of evidence from the study of idioms, that "just as there is evidence that a string generally agreed to be formulaic may or may not have a high frequency in even the largest of corpora, so it is also not possible to assert that all frequent strings are prefabricated" (ibid.: 31). However, Wray (2009) admits that frequency criteria may be useful and theoretically justifiable in practical terms: "even for those who question whether frequency determines formulaicity, it would take a strong case to successfully argue that the frequent examples of formulaic sequences are not a good place to start" (Wray 2009: 102). Wray draws attention to her prime criteria: formulaicity or prefabrication, which is, in comparison, wholly absent from Gries' definition, and may be said to belong to a different area of definition. It seems that whereas Wray's intention is to introduce a 'coverall' term and a definition applicable to many diverging ideas of what language patterning is, Gries' approach is of a more pragmatic kind, one tailored for work within corpus linguistics, outlining clear conditions which are possible to detect when studying authentic stretches of text. Prefabrication and holistic storage are not, on the other hand, features which can be scientifically identified this way, and thus Wray's definition may come across as too vague to be of any use in text-based studies. However, it seems as if these more abstract features should serve as a backdrop for any other definition, as they tie them to broader theories on how we acquire, store, and employ language patterns, and create an aim for studies of language patterning that goes beyond the identification of specific forms. Erman and Warren (2000) seem to take on such a stance in their study of 'prefabs', in which they seek to investigate "the impact

that prefabricated language has on the structure of a text and of the effort involved in encoding and decoding it" (Erman and Warren 2000: 30). However, their definition of 'prefab' makes no explicit reference to the notion of 'prefabrication', which must be said to be suggested by the term itself:

> "A prefab is a combination of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalisation" (ibid.: 31).

This definition echoes Pawley & Syder's (1983) discussion of 'nativelike selection', and gives prominence to an aspect of language patterning relevant to the study of interlanguage, but seems to lack both the specific criteria prompted by Gries, as well as Wray's reference to possible holistic representation in the mental lexicon. Such a reference might be inherent in the term 'conventionalisation' in the definition, but, as admitted by the authors, conventional language does not strictly imply prefabrication as opposed to grammatical 'on-line' generation (Erman and Warren 2000: 33), and may otherwise be hard to detect in terms of individual representation and opinions on what is 'conventional' in their language. John Sinclair's (1991) well-known definition of the 'idiom principle' is comparable to Wray's definition of a formulaic sequence, and similarly draws attention to the way our mental lexicon stores and selects word-combinations:

> "the principle of idiom is that a language user has available to him a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (Sinclair 1991: 110).

The idiom principle is juxtaposed and co-existent with an 'open choice' principle, similar in design to the foundations of generative grammar, where "the only restraint is grammaticalness" (ibid.: 109). With these two principles, Sinclair paves the way for a conception of language as operating within a rich memory system, where word-combinations can be stored as sequences and as individual words simultaneously: "Just because a multi-word expression is stored and processed as a chunk does not mean that it does not have internal structure" (Bybee 2010: 36). This postulation of a rich memory system rather than one free from redundancy holds further implications for our conception of language patterning, as it becomes easier to imagine other kinds of information being stored in such a system, including "phonetic detail for words and phrases, contexts of use, meanings and inferences associated with utterances" (ibid.:

7). Gries indirectly refers to this feature in his definition of 'phraseologisms', in that he does not require them to be semantically non-compositional. The semantic meaning of a phraseologism may, in Gries' view, be deduced on the basis of the individual semantic properties of its component parts (Gries 2008: 6), which should imply that they be stored in our memory systems along with the phraseologism itself. On the surface, Gries' definition seems to be less theoretically assuming and thus more suitable for a text-based analysis which seeks to arrive at well-founded conclusions, but the additional assumptions about prefabrication could easily be included in such a definition in a more or less explicit fashion, without endangering the validity of conclusions reached in a study of phraseologisms. The fact that other definitions go further in their speculations regarding this question is undoubtedly due to the promising answers such speculations may provide.

Even though it might be tempting to leave the difficult question of holistic storage and what is often referred to as 'psycholinguistic validity' (cf. Schmitt et al. 2004) to the domain of psycholinguistic experiments or mentalist theories, it seems inevitable to also consider it in studies of text, if we are to take advantage of some of the interesting implications a prefabricated status might entail, as provided by e.g. a cognitive linguistic framework. This fact is also increasingly acknowledged by linguists working with corpora, e.g. De Cock (2004), who conclude that "the psycholinguistic validity of automatically extracted recurrent sequences of words and the relationship between recurrence and the storage of sequences of all kinds as wholes in the brain (...) will need to be dealt with in greater detail in the near future" (De Cock 2004: 243). Among other features, prefabricated language patterns "allow the language user to be more fluent while at the same time freeing up cognitive resources for other language processes" (Schmitt et al. 2004: 128). This ease of processing should be particularly valued by learners of a second language, who will benefit from any such widening of cognitive 'work-space' for particularly demanding processes such as correct lexical retrieval. In addition, processes of prefabrication in combination with high frequency of use are believed to explain language change of phonetic, semantic and pragmatic nature (Bybee 2010: 48-49), and these are important processes particularly for understanding how recurrent word-combinations might contribute to traces of non-nativeness found in interlanguage corpora.

2.2.2.3 Phraseological terminology employed in the present thesis

It thus seems that even though most researchers working with the phenomenon of language patterning acknowledge the importance of the possibility of cognitive prefabrication, prefabricated sequences are hard to detect from studies of authentic text other than on the basis of intuitive ideas of conventionalisation. In addition, prefabrication cannot be regarded as a static feature of a word-combination, but rather as processes of storage and retrieval which "differ from one individual to another, and can differ from one time to another for the same individual depending on a wide range of factors such as changes in proficiency, changes in processing demands, and changes in communicative purpose" (Read and Nation 2004: 25). Agreeing with Sinclair's view of the position of intuition in corpus methodology, that "once a description arrived at with maximum objectivity has been achieved, the intuitions and responses of the human researcher are essential for interpretation of the phenomena" (Sinclair 1999: 1), a definition according to Gries' parameters seems to be appropriate as a working definition of formulaic sequences for the present thesis, albeit with the inclusion of Wray's emphasizing but non-conclusive take on prefabrication: "is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use" (2002: 9). I also adopt the term 'formulaic sequences' from Wray (2002), as it conveys the sequential nature of word-combinations as well as a notion of formulaicity, which I believe to be more theoretically objective in nature than the notion of prefabrication or conventionality. The term is also fairly transparent, making it more generally accessible than Gries' 'phraseologism', a term which echoes the specific field of phraseology. Additional terms will be employed where appropriate in reference to studies which make use of these other terms, but the terms 'formulaic sequence' and 'recurrent word-combination' will be used predominantly in the analysis, where the former is subject to identification according to Gries' parameters and the latter is an all-comprising term for n-gram search results, defined solely on the basis of frequency of consecutive lexical co-occurrence. Erman and Warren's emphasis on the nativelike nature of formulaic sequences will also be discussed, but it will not be seen as a defining feature, mainly because it seems plausible from the perspective of interlanguage studies to consider occurrences of recurrent word-

combinations found in the interlanguage data as formulaic, even if they intuitively and on the basis of comparison with native language data are considered non-nativelike.

It is not the intention of this thesis to detect *all* instances of formulaic sequences in the selected corpora, nor is it aiming to make far-reaching conclusions on the question of prefabrication and holistic storage, or the general nature of formulaicity. Instead, I will assume that empirical evidence of high frequency, though not necessarily an exclusively defining or even necessary feature of formulaic sequences, suggests mental entrenchment in such a way as to make the word-combinations eligible for further analysis (cf. Altenberg 1998, DeCock 2008). In addition, in a comparison between interlanguage and native speaker data diverging patterns of significant frequency may be of interest in itself, as possible contributors of 'non-nativeness' in a learner text. Thus I will agree with Gries and require that the frequency of the word-combinations examined in the study is 'larger than expected on the basis of chance', employing statistical tests where appropriate to substantiate this claim. Furthermore, the word-combinations discussed will predominantly be continuous and based on form rather than lemmas, as a further consequence of the method employed where an initial n-gram search retrieves continuous strings of identical form. Similarly, cognitive constructions with few or no searchable lexical items such as '*What* BE SUBJECT *doing* Y' (Bybee 2010: 77), although theoretically interesting, will not be included here. It is however useful to consider what Erman and Warren (2000) terms 'restricted exchangeability', as a test for the semantic/pragmatic unity and degree of formulaicity of the word-combination, which implies that "at least one member of the prefab cannot be replaced by a synonymous item without causing a change of meaning or function and/or idiomaticity" (Erman and Warren 2000: 32). In the context of comparative study of native and interlanguage, the lexically *continuous* retrievals an n-gram search will provide are considered interesting, as features of form that differ across the two corpora may be indicators of non-nativeness or notable interlanguage behaviour, and, as mentioned above, identical word-combinations that differ according to features of e.g. frequency, distribution or function may also be worthy of analysis. While acknowledging that formulaic language may be inherently variable (cf. Sinclair 1999b), this contrastive approach thus concentrates on usage

patterns and possible prefabrication of various *fixed* word-combinations, in the search for interlanguage features.

Schmitt and Carter (2004) conclude in their review on formulaic sequences that "if creatively-generated language was cognitively more efficient, we would not expect to find formulaic sequences realizing functional language usage nearly as frequently as we do in corpus evidence" (Schmitt and Carter 2004: 5), and thus point to the fact that formulaic language can be seen to perform certain *functions* in language use, in addition to the overall cognitive benefits. Native language and interlanguage users may need to express different functional types, or attempt to realize similar functions by the use of patterns that differ in formal qualities, and I will in the analysis and conclusion make use of different functional classifications of recurrent word-combinations, e.g. those presented in Altenberg (1998), as well as discussing possible extra-linguistic incentives for performing certain functions in the two language populations, as outlined in e.g. Wray (2002). Within a functional perspective, one might even assume that formulaic language appears as a result of the continuous performance of particular functions in a language community, and the conventionalising forces that make certain forms accepted as the common, or native-like, alternative for such functional performance. Schmitt and Carter further argue that "there seems to be a link between the need and desire to interact and the use of formulaic sequences" (Schmitt and Carter 2004: 11), and the pragmatic aspect of formulaic language will be a prominent part of the analysis, hence the focus on semantic/pragmatic unity in the definition. Following the above discussion, the working definition of a formulaic sequence employed in the forthcoming study is presented below:

> A formulaic sequence is defined as the co-occurrence of two or more consecutive lexical items which functions as one semantic/pragmatic unit in a clause or sentence, which is, or appears to be, prefabricated and conventionalised, and whose frequency of occurrence is larger than expected on the basis of chance.

The functional perspective and the notion of semantic/pragmatic unity are perhaps the aspects most open for discussion in this definition, and the exact properties of these parameters should thus be subject to modification and discussion throughout the analysis. As mentioned above in relation to Erman and Warren's (2000) definition based on nativelike conventionalisation, viewing learner populations as less unified

than native speaker populations may bring the whole notion of speaker unity and conventionalisation into question in an interlanguage analysis. A significantly high frequency level may serve as a regulatory feature, ensuring some indication of formulaicity beyond intuitive judgements, but there are also a number of aspects which tie groups of second language learners together by virtue of being learners. Theories presuming universal cognitive processes for first and second language production might serve to explain the appearance of certain recurrent patterns on the basis of other factors besides conventionalisation. These processes may also explain evidence of transfer of native language features in the interlanguage output.

2.2.3 The learner, Interlanguage and Evidence of Interlanguage Formulaicity

The term 'interlanguage' was coined by Selinker (1972), and is usually used to refer to, in Smith's (1994) words: "the systematic linguistic behaviour of learners of a second or other language" (Smith 1994: 7). The 'inter' in 'interlanguage' suggests that the language it is used to describe is somehow at an intermediate stage, and one might say that a learner's interlanguage is indeed a reflection of his or her move towards nativelike language behaviour. However, although a language learner commonly seeks to obtain command of a certain set of native speaker norms, it is considered important not to view learners as "just weak imitations of native speakers" (Wray 2002: 195), or instances of interlanguage as "merely imperfect reflections of some norm" (Smith 1994: 7), but rather consider the language produced by learners as "possessing systematic features which can be studied in their own right" (ibid.). It is for these reasons that the more transparent term 'learner language' is employed as the main term in the forthcoming analysis.

This view is particularly important regarding the study of formulaic language, as it makes us able to draw lines between the processes of patterning that take place in native language and the evidence we find of formulaicity in interlanguage data. Theories of the cognitive processes underlying native and second language production are many diverging (cf. e.g. Abrahamsson and Hyltenstam 2009), but theories proposing that similar or even identical processes operate in both languages hold interesting implications for studies of patterns in second language production. Contrasting with research that assume first language acquisition and second language

acquisition to be inherently different, Barlow (1996), among other linguists within the cognitive usage-based tradition, propose that "the underlying cognitive terrain is essentially language-independent and, thus, is as suitable for the second language as for the first" (Barlow 1996: 27). In a similar fashion, Bybee (2010), discussing diachronic and synchronic language change, believes that our cognitive system is "a complex adaptive system" (Bybee 2010: 2), with domain-general processes that function in combination with language use to change and develop language in a continuous fashion. These views suggest that what happens in a second language and in second language acquisition is not inherently different from what happens in native language and native language acquisition, which in turn makes it possible to apply our knowledge of e.g. language patterning in native language production to our accounts of interlanguage findings.

In contrast, Kjellmer (1991) proposes a second language acquisition model suggesting that whereas native speakers process (spoken) language predominantly according to Sinclair's idiom principle (see section 2.2.2.2), second language learners build up their utterances on the basis of individual building blocks, according to the principle of open-choice:

> "While the typical moderately fluent native speaker makes considerable hesitation pauses between often quite long sequences of words (…), the typical moderately fluent learner pauses after every two or three words. It seems reasonable to believe that the difference between them in this regard can be ascribed largely to a difference in the automation of collocations" (Kjellmer 1991: 124).

The assumption that learners 'over-analyse' input in such a way would explain e.g. observations of learner language which is "meaningful but not nativelike" (Wray 2009: 20), e.g. *take advantages of*, *on the meantime* (Wray 2002: 1999). Similarly, Götz and Schilk (2011) believe that "child language acquisition is by its very nature more holistic than analytic" (Götz and Schlik 2011: 83), whereas "learners create meaning by combining individual words, possibly without the awareness of additional meanings of multi-word sequences" (ibid.: 85). However, this behaviour is not unanimously ascribed to radically different processing differences between learners and native speakers, or between languages learnt as a child and as a teenager or adult. Theories supporting a critical period for language learning, or "a sharply defined age of termination, after which normal development is no longer possible" (MacWhinney 2009: 346), suggests that the processes that create our knowledge of our first language

is no longer available to us beyond a certain (early) age, and that second language acquisition processes are somehow inherently different. MacWhinney (2009) believes language acquisition to be best explained in terms of a Unified Model of first (L1) and second (L2) language acquisition, which "views age-related changes in L2 learning in terms of entrenchment, competition, and transfer, rather than the expiration of a critical period" (MacWhinney 2009: 363). Entrenchment of native language categories and lexical inventory in adult learners may lead to competition between these and the properties of the second language, which in turn may result in problems of transfer from the native language. Transfer may be evident on a more subtle level in the case of the recurrent word-combinations produced by advanced learners, where there is commonly a non-nativelike *pattern of usage* rather than a non-nativelike form (cf. Ringbom 1998: 49), such as the overuse and underuse of certain word-combinations, as well as semantic and pragmatic differences. Such subtle differences may also relate to processes that sometimes occur with highly frequent word-combinations in native speech, which I will get back to in the analysis of the particular word-combinations in chapters 4 and 5. When considering language as "an embodied activity that occurs in real time, in real situations and passes through real cognitive systems" (Bybee 2010: 221), we may rather explain the difficulties experienced by second language learners in terms of these continuous processes, in addition to external factors relating to the language learning situation.

Following this, the solution to problems of transfer, in the all-comprising sense outlined above, would, according to Bybee (2009) require "sufficient exposure to the categories of the L2" (Bybee 2009: 233), in order to 'override' entrenched native language categories, forms and usage patterns. Concerning formulaic sequences, it also seems intuitively plausible that sufficient exposure should be one of the key elements required for a learner to memorize an adequate portion of the formulaic inventory of a target language. In addition, to ensure a native-like usage pattern, the level of exposure needs to allow for aspects of linguistic context, function and semantic/pragmatic meaning to be stored alongside the sequence in the rich memory system. In addition to explanations referring to our cognitive inventory, Wray (2002) refers to more external forces which may explain the linguistic behaviour of adult learners, particularly in relation to formulaic sequences:

> "Formulaic sequences are selected as a response to specific needs in communication and processing. Precisely which formulaic sequences different second language learners use should depend on their different priorities, and on the situations in which they find themselves. Being most prevalent where the need is greatest, the formulaic sequences produced by any individual should correlate with an independent assessment of his or her socio-interactional and processing priorities at that time" (Wray 2002: 144).

The impression of a critical age is, according to Wray, related to these differences in priorities:

> "The critical age (...) is a conglomeration of factors which affect the individual's approach to learning. The learning itself is subservient to the real agenda, which is to accommodate the immediate needs – all of them – of the individual, not only as a learner but as a functional entity in his or her own complex world" (Wray 2002: 213).

Evidently, specific needs of individual language users can be hard to identify, particularly through a corpus study of the language of a large number of individuals, but since details are provided about the situational context of the subjects and the recording process (see also chapter 3 on material method), these details, in addition to the linguistic context of word-combinations and general impressions of the language learner situation, are useful considerations in the attempt to account for usage patterns in both native language and interlanguage. This study of recurrent word-combinations in learner language builds on the assumption that there is something common to 'the language' of learners, and that there are similarities between the language output of learners in general as well as learners from a specific mother tongue background, which can be revealed through corpus analysis of this output. Wray (2009) debates whether such a unity can be ascribed to language production in at all, with particular reference to the production of formulaic sequences: "Formulaicity is viewed as the property of a particular string as it is handled by a particular individual" (Wray 2009: 11). However, even though "what is formulaic for one person need not be formulaic for another" (ibid.), there seems to be both practical and theoretical reasons for viewing tendencies found in language data as valuable contributions to our understanding of formulaic language, and this is also acknowledged by Wray: "even an account based on the individual's knowledge will recognize that many word strings are likely to be formulaic for most native speakers - that is what it means to know the same language" (ibid.). Even though it is likely that there is greater variation in learner language than in native language in terms of

formulaicity, it should thus be possible to draw conclusions on learner language the basis of a representative data selection (see section 3.2).

2.2.4 Recurrent Word-Combinations and Formulaic Sequences in an Interlanguage Perspective: Previous Research

The considerations outlined in the previous section agree with Ellis and Robinson's (2009) account of language, communication, and cognition as being "mutually inextricable" (Ellis and Robinson 2009: 3), and propose that we should consider a multitude of factors in the analysis of authentic text. This also seems to be the stance of many researchers working with recurrent word-combinations, as reflected in the various definitions discussed above. Practical concerns have, as mentioned in section 2.1, long complicated the compilation of corpora of spoken language, but recent years have seen an increase in corpus-based studies of speech patterns. Compiling a corpus of spoken *learner* data involves additional considerations and difficulties, which is further discussed in the next chapter, and the majority of studies on formulaicity in learner language have thus so far been predominantly based on written language data. This is one of the reasons why it is important to conduct studies on spoken learner language in general. The following presentation on previous research is only a brief selection of relevant articles which approach formulaicity in spoken- and learner language, with focus on the studies that have inspired the forthcoming analysis.

Several studies have set out to test Kjellmer's (1991) preliminary hypothesis about the differences between language processing in native and learner language (section 2.2.3) through the analysis of learner corpora, gaining knowledge about the general nature of formulaicity in interlanguage along the way. In one of these studies, De Cock et al. find, perhaps not surprisingly, that "learners do use prefabs" (De Cock et al. 1998: 72), and De Cock later concludes that "Kjellmer's assumption that learners' building material is individual bricks rather than prefabricated sections appears to be simplistic" (De Cock 2004: 243). Such findings further suggest that a unified model of language acquisition is reasonable, and that language patterning is at work in second language acquisition processes as it is in native language processing, albeit in a fashion which requires increased input and correction in order to achieve nativelike competence. DeCock et al. (1998) concludes that even though learners, contrary to

Kjellmer's prediction, seem to make use of the idiom principle, the formulaic sequences found in the interlanguage data "(1) are not used with the same frequency, (2) have different syntactic uses, and (3) fulfil different pragmatic functions" (De Cock et al. 1998: 78), thus displaying a non-nativelike *usage pattern* rather than necessarily employing non-nativelike constructions of form. Granger (1998a) reaches a similar conclusion in her study of 'formulae' in written learner language, arguing that her findings of overuse of certain word-combinations shows a tendency for learners to "'cling on' to certain fixed phrases and expressions which they feel confident in using" (Granger 1998a: 156), and that "while the foreign-soundingness of learners' production has generally been related to the *lack* of prefabs, it can also be due to an excessive use of them" (ibid.: 155). Such 'phrasal teddy bears' have been reported in several studies, and may play a major role as contributors to an impression of non-nativelike language production in interlanguage data that otherwise show few overt traces of non-nativeness. These frequently used word-combinations will evidently be easier to identify in a frequency search than non-nativelike combinations that are present, but not frequent, but in a search for non-nativeness it is useful to consider the mere frequency of these patterns as well as their possible extended pragmatic use as compared to native speech. Very frequent patterns that deviate from a native norm are likely to be undesirable markers of non-nativeness in learner speech because of their pervasiveness and because they may indicate a noticeable underuse of other, nativelike patterns. These are findings that go beyond the capacity of intuitive identification, and are difficult to detect without the use of corpora and corpus methods.

In his paper "On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations" (1998), Bengt Altenberg makes use of a corpus of spoken native English (the London-Lund corpus) and what has later been termed "the corpus driven 'recurrent word combination' method" (De Cock 2004: 227), in order to investigate word-combinations in the spoken language of native speakers of English, as a part of the project presented in Altenberg and Eeg-Olofsson (1990). Altenberg is particularly interested in the evidence he finds for "(…) the pervasive and varied character of conventionalized language in spoken discourse" (Altenberg 1998:120), and further concludes that "conventionalization of complete sentences or clauses is

mainly a pragmatic phenomenon: certain expressions are needed to convey various recurrent speech acts and discourse strategies and many of these are conventionalized by frequent use" (ibid.:121). Making use of Altenberg's method, which will be described further in the following chapters, De Cock (2004) investigates spoken learner language, and shows similar findings from the French component of the LINDSEI corpus, while she also finds that, in comparison with the native speaker corpus LOCNEC, "advanced learners' use of frequently recurring sequences of words displays a complex picture of overuse, underuse, misuse of target language NS sequences and use of learner idiosyncratic sequences" (De Cock 2004: 243). De Cock highlights the importance of contrastive studies of formulaicity in learner language for pedagogical theory and application, since "not only do they provide us with real NS usage, but they also bring to light the sequences learners appear to find problematic" (ibid.). This fact is also stressed by Granger in relation to English language teaching (ELT) (1998b):

> "It is paradoxical that although it is claimed that ELT materials should be based on solid, corpus-based description of native English, materials designers are content with a very fuzzy, intuitive, non-corpus based view of the needs of an archetypal learner. There is no doubt that the efficiency of EFL tools could be improved of materials designers had access not only to authentic learner data, with the NS (native speaker) data giving information about what is typical in English, and the NNS (non-native speaker) data highlighting what is difficult for learners in general and for specific groups of learners" (Granger 1998b: 7).

Granger (1998a) finds in a study of written English produced by native speakers of French, that advanced learners produce "too few native-like prefabs and too many foreign-sounding ones" (Granger 1998a: 158), and claim that we need better descriptions of recurrent patterns in both target language, native language and non-native language in order to achieve a better understanding of language acquisition processes, and the distribution of certain language patterns. A good description of learner language is needed "because not all learner problems are transfer related" (ibid.), and following this, it seems essential to also conduct studies of learner language in different situational contexts, so as to shed light on contextual factors which might have an impact on e.g. the occurrence of recurrent and/or prefabricated language. In addition, it is important to map out differences between different mother tongue populations, which is made possible through the compilation of multinational corpora such as LINDSEI. So far, not many studies have been conducted based on the

Swedish component of LINDSEI[2] or, naturally, the (unfinished) Norwegian component, and it is thus particularly interesting to investigate recurrent word-combinations in these two corpora.

De Cock underlines the connection between studies of recurrent word-combinations and the question of psycholinguistic validity: "does recurrence actually cause a sequence to be stored as a unit or is recurrence a result of a sequence being stored whole and therefore easily accessible?" (De Cock 2004: 243-244), also claiming that "SLA research would gain a great deal from a better understanding of phraseology, in general, and formulae, in particular" (De Cock 1998: 76). Dahlmann and Adolphs (2009), in their study of the word-combination *I think* in the English Native Speaker Interview Corpus (ENSIC), claim that a 'multimodal approach', with attention to e.g. pauses and gestures, must be employed to corpus-derived findings, if we are to make more informed suggestions about holistic storage, which is "almost impossible to measure [...] directly" (Dahlmann and Adolphs 2009: 125-126). Similarly, Lin and Adoplhs (2009) explore, through a corpus of spoken learner language, the notion of phonological coherence as a criterion for identifying prefabricated language, based on the assumption that "if phraseological units are always phonologically coherent, phonological coherence might be established as an alternative to the psycholinguistic methods currently adopted to explore the reality of holistic storage and processing of phraseological units" (Lin and Adoplhs 2009: 35). They further argue that learner corpora are particularly well-suited for uncovering phonologically coherent word-combinations, since learner speech is typically slower and more hesitant, making formulaic language stand out in terms of "fluency and specific phonological features" (ibid.: 39). Since the LINDSEI corpus is tagged for filled and unfilled pauses, Brand and Götz (2011) are able to show through a study of the German subcorpus, that German advanced learners of English significantly overuse unfilled pauses, compared to the native-speaker mean extracted from the LOCNEC corpus (Brand and Götz 2011: 263), and it is possible to assume that these differences are due to the non-

[2] cf. http://www.sprak.gu.se/forskning/forskningsomraden/korpuslingvistik/korpusar-vid-spl/swe-lindsei/ [visited 17.10.2011].

native speakers' considerable planning constraints, but also to a smaller or less established inventory of holistically stored formulaic sequences. These studies suggest that there is a close relation between fluency and formulaicity in spoken learner language.

Since the forthcoming analysis is predominantly based on Altenberg (1998) and De Cock (1998; 2004), I will discuss their methods, findings, and classifications of recurrent word-combinations in further detail in the following chapters. I will also make reference to Karin Aijmer's research on material from the Swedish component of LINDSEI (2004; 2009; 2011), and other studies relevant to specific word-combinations or their surrounding features.

2.3 Recurrent Word-Combinations in Spoken Learner Language: Summary

It seems clear on the basis of the theory presented in this chapter that more research is needed to investigate the part recurrent word-combinations play in spoken interaction, and how their occurrence, absence or non-native usage patterns affects our impression of learner language. Through identifying formulaic language in learner speech and mapping out their patterns of use, we may also understand more about how they influence second language acquisition and production. Studying the speech of learners from different mother tongue backgrounds may give us valuable information on the impact previous language experience or cultural factors may have on the production of formulaic sequences. The LINDSEI corpora seem to be very useful in this respect, as they consist of learner speech data which is controlled for many factors that would otherwise affect production. It is however important to evaluate this data before any conclusions can be made on the basis of findings from it, and chapter 3 aims to determine the validity of the corpora for the analysis of recurrent word-combinations and formulaic language. The next chapter will also briefly describe methodological concerns prior to the analysis.

3 Material and Method

3.1 Introduction: A Study of Empirical Data

As in any other empirically based study, the initial choice of material in a corpus study is crucial for the validity and scope of the results it provides. Setting out to describe and explain the occurrence of recurrent word-combinations in learner language, several considerations need to be addressed in order to assure that the chosen material and the method employed are appropriate for providing suitable answers. According to Granger (1998b), second language acquisition research in general has as its main goal to "uncover the principles that govern the process of learning a foreign/second[3] language", and this process "is mental and therefore not directly observable, it has to be accessed via the product, i.e. learner performance data[4]" (Granger 1998b: 4). Similarly, Ellis and Barkhuizen (2005) assert that "all researchers who accept the primacy of learner language as data for investigating L2 acquisition accept that learners' use of the L2 in some way reflects their L2 competence/proficiency" (Ellis and Barkhuizen 2005: 364). As stressed in chapter 2, evidence from *authentic* material is the foundation of many linguistic studies today, which emphasise that "it is important to base one's analysis of language on real data – actual instances of speech or writing – rather than data that are contrived or 'made-up'" (Meyer 2002: xiii). Once this view is established, however, authentic material in its purest sense, "language that is situationally and interactionally authentic" (Ellis and Barkhuizen 2005: 7), may for various reasons be difficult to collect for research purposes. Section 3.2 below will discuss authenticity in terms of the compilation of spoken learner corpora, as well as representativeness, which also determines the

[3] Some researchers postulate a distinction between 'foreign language acquisition' (FL) and 'second language acquisition' (SL) (i.e. Granger 1998b), according to whether the language in question is learnt in a foreign environment or in an environment where the language is spoken. Since this distinction is not directly relevant for the present study, and since there does not seem to be a general consensus on terminology among different researchers, the term 'second language' will be used to refer to English learnt in a foreign environment, unless otherwise stated.

[4] Certain methods are of course available for observing the effects of language behaviour in the brain, such as functional magnetic resonance imaging (fMRI) (see e.g. Gernsbacher and Kaschak 2003), but, as e.g. Ellis and Barkhuizen (2005) notes, the knowledge we can gain from these experiments are still marginal, and "by and large, researchers are forced to infer competence from some kind of performance" (p. 6).

validity of corpus results. Every language situation includes a variety of variables connected to their subjects and settings, and these variables need to be accounted for if results are to be generalized to a broader language population, as well as to allow for replication and comparison of studies. This has been a problem in e.g. applied linguistics, where important issues have suffered because "researchers have not been comparing like with like" (Granger 1994: 44).

Agreeing with the importance of authenticity and representativeness, Gries (2009), as seen in chapter 1, includes authenticity in his definition of a corpus, describing it as "a machine-readable collection of (spoken or written) texts that were produced in a natural communicative setting, and the collection of these texts is compiled with the intention (i) to be representative and balanced with respect to a particular linguistic variety or register or genre and (ii) to be analyzed linguistically" (Gries 2009: 411). In corpus linguistics, the careful selection and compilation of material, component (i) in Gries' definition, is often not performed by the analysts themselves, as the typical electronic form and (more or less) unrestricted availability generally allows for a joint investigation of the same material by several researches. While this co-operation saves time and enables close scrutiny of both the material and the various research results arising from it, it also leaves the analysts responsible for learning the build-up of the corpus in question, and its strengths and weaknesses in terms of particular research questions: "(…) because it is virtually impossible for the creators of corpora to anticipate what their corpora will ultimately be used for, it is also the responsibility of the corpus user to make sure that the corpus he or she plans to conduct a linguistic analysis of is valid for the particular analysis being conducted" (Meyer 2002: 53). According to Ellis and Barkhuizen (2005), "the key methodological issue is what kind of performance provides the most valid and reliable information about competence" (Ellis and Barkhuizen 2005: 21).

Even though the nature of particular corpora may restrict the number of methods available to investigate it, an electronic corpus generally offers many and diverse possibilities in terms of analysis. As Gries (2010b) observes, "branches of linguistics that have been using corpora or text databases have always been among the most quantitatively oriented subdisciplines of the field" (Gries 2010b: 5), and frequency counts of words or word-combinations are at the centre of most corpus studies. The

present study is no exception, and a quantitative approach will be adopted to inquire into the appearance of recurrent word-combinations in learner speech. Based on Granger's (1994) 'Contrastive Interlanguage Analysis', the study is contrastive in nature, and section 3.3.2 will discuss the contrastive method as well as the quantitative methodology employed in e.g. Altenberg (1998). As mentioned in chapter 2, the qualitative part of the analysis will include an identification of formulaicity and functionality, building on the discussion of formulaicity in section 2.2.2.

3.2 Material

> "One of the fundamental difficulties with researching how people learn languages is separating out the many interacting variables that operate in realistic conditions of language use. The most scientifically robust way to observe language learning and performance would be to put people into an artificially manipulated situation in which the causes of their behaviour could be tracked and attributed. Yet their use of language would then also be artificial, and may not reflect what they would do in the real world" (Wray 2009: 153).

As mentioned above, a corpus should ideally consist of texts that are somehow authentic, as well as representative of the language population whose language behaviour one is interested in. The validity of results from corpus analyses is thus partly determined by the corpus material, particularly in terms of its authenticity and representativeness. Granger (2008) argues for strict design criteria for learner corpora, since "learner language is influenced by a wide variety of linguistic, situational and psycholinguistic factors, and failure to control for these factors greatly limits the reliability of findings in learner language research" (Granger 2008: 263). Furthermore, Granger (1998b) stresses that "it is especially important to have clear design criteria in the case of learner language, which is a very heterogeneous variety: there are many different types of learners and learning situations" (Granger 1998b: 7). This focus on situational variables is a reflection of the special attention paid to the wider context in corpus studies, and an appreciation of the importance of these variables for explaining features of language performance. Granger (2008) divides learner corpus design criteria into two sets of variables, one pertaining to the learner and the other to the 'task', or situation:

Learner variables	Task variables
Age	Medium
Gender	Field
Region	Genre/text type
Mother tongue	Task type
Learning context	Conditions
Proficiency level	
L2 exposure	
Other FL	

Table 3.1: *Learner corpus design criteria (table adapted from Granger 2008: 264)*

These variables are especially important to account for in a contrastive analysis, as a contrast may only provide interesting results if we know what variables differ between the corpora we are comparing. It seems that what Granger terms 'task variables' are the most important characteristics of a corpus for determining its authenticity in terms of similarity to real-life situations. I will discuss these task variables and their relation to the concept of authenticity in section 3.2.1. 'Learner variables', on the other hand, are important as determinants for generalization and representativeness, and I will discuss how LINDSEI, LOCNEC and the LINDSEI subcorpora may be said to be representative of broader language populations in section 3.2.2.

3.2.1 Task Variables and Authenticity as a Measure of Validity

Ellis and Barkhuizen (2005) argue that "in any study it is necessary to *demonstrate* the validity of the data that have been collected", and this consideration is perhaps of special importance in a corpus study, because of the general belief among most corpus linguists in the supremacy of authentic data for descriptive analyses as well as the possibility for "access to the language-related capabilities of the mind" (Chafe 1992: 88) through authentic speech and writing. Ellis and Barkhuizen further claim that in the collection of learner data, the construct validity, i.e. the extent to which a study is measuring what it set out to measure, "is best established by demonstrating that the performance it taps reflects, as far as possible, the kind of use for which language is designed and acquired" (Ellis and Barkhuizen 2005: 21). Researchers who conduct experiments and elicitation procedures may be able to defend the validity of

their data, but, as Chafe (1992) observes, "the unnaturalness of the procedure can be quite troubling" (Chafe 1992: 85). Similarly, Granger (1998b), commenting on the use of elicitation in second language acquisition research, notes that: "The artificiality of an experimental language situation may lead learners to produce language which differs widely from the type of language they would use naturally" (Granger 1998b: 5), in accordance with e.g. the Observer's Paradox. In order to reach valid conclusions on natural language use, then, it seems as if the most appropriate approach is "the observation of naturally occurring overt behaviour" (Chafe 1992: 88), i.e. corpora of authentic text.

However, authenticity may pose a particular problem for compilers of all kinds of learner corpora, as learners "rarely use the target language to go about their normal business" (Granger 2008: 261). A typical situation for the language learner may rather be some sort of test situation in connection with his or her language education, and indeed, any situation where the learner speaks his second language will typically lead to a higher level of language-consciousness than situations where he speaks his mother tongue. Granger thus proposes a "naturalness continuum" (ibid.: 261) for learner language, with informal interviews or free compositions naturally ranking higher than e.g. reading aloud (ibid.). The term 'learner corpora', Granger believes, should be restricted to continuous, contextualized texts, which "allow learners to choose their own wording rather than being requested to produce a particular word or structure" (ibid.), thus excluding pure elicitation, but still allowing for language produced with some externally imposed direction. Ellis and Barkhuizen (2005) argue in favour of authentic learner data, agreeing that "what counts in the study of interlanguage development is learners' procedural knowledge; and this can only be investigated by means of naturally occurring data" (Ellis and Barkhuizen 2005: 48). However, they take on a similar stance to Granger (2008), by acknowledging that in actual research, only relying on naturally occurring data "is probably too extreme and is certainly impractical" (Ellis and Barkhuizen 2005: 48). In agreement with this position, they postulate a distinction between (1) naturally occurring samples, (2) clinically elicited samples and (3) experimentally elicited samples, with clinically elicited samples taking on a middle position, where "some control is exercised through the choice of task but learners are expected to be primarily engaged in

message conveyance for a pragmatic purpose, as in naturally occurring language use" (ibid.: 24). The 'pragmatic purpose' of the clinically elicited texts should thus create an environment where the language learner is less aware of her own language performance, so that the performance is as implicit (procedural) as possible, rather than explicit (declarative), and in this respect closer to natural language production than experimentally elicited samples. Clinically elicited texts may also have advantages over naturally occurring texts in terms of the increased control of variables this format entails.

3.2.1.1 Task variables: LINDSEI and LOCNEC

Granger (2008) postulates, as seen in table 3.1, five task variables to consider in learner corpus compilation and analysis: *Medium*, *field, genre/text type, task type* and *conditions*. In an assessment of authenticity the task type and conditions are perhaps of most relevance, though it is ultimately the combination of variables which makes a situation appear natural or not. The Louvain International Database of Spoken English Interlanguage (LINDSEI) is a corpus consisting of spoken informal interviews of learners of English, divided into subcorpora according to the mother tongue of the subjects interviewed (Gilquin et al. 2010). The Swedish subcorpus was compiled at the University of Gothenburg in Sweden, and the Norwegian interviews are carried out at Hedmark University College in Norway. The comparable corpus of native speaker English, the Louvain Corpus of Native English Conversation (LOCNEC), is built up according to the same principles as LINDSEI (ibid.), and the interviews were carried out at Lancaster University, UK. The interview sessions in LINDSEI and LOCNEC are recorded non-surreptitiously and with provided guidelines as to the topics of conversation (typically topics involving the subjects personally), including follow-up questions from the interviewer, and a picture description task towards the end of the interview (see Appendix). There is no set time limit, and the interview is not used for any sort of external assessment of the subjects, who are all university students majoring in English.

The resulting LINDSEI and LOCNEC texts may thus be said to belong to what Ellis and Barkhuizen (2005) terms clinically elicited *general* samples, which "can lay claim to 'some sort of relationship with the real world' in that they involve the kind of

communicative processes involved in the real-world" (Ellis and Barkhuizen 2005: 31). As opposed to experimentally elicited samples, these samples consist of texts which aim to achieve a certain communicative task, such as, in the example of LINDSEI and LOCNEC, telling a story, answering a question or describing a picture. Even though the learner texts are to a certain extent controlled thematically by the previously set topics, the task may be considered linguistically 'open', since subjects are allowed to choose linguistic form, and since extracting certain grammatical features is not the primary aim of the collecting of data.

The interview situation in LINDSEI and LOCNEC is thus in many respects similar to communicative situations in 'the real world', but the interviews "were not produced for real communicative purposes, but for classroom (and corpus collection) purposes" (Gilquin and De Cock 2011: 157), and this in addition to the surveillance of the tape-recorder may have an effect on the naturalness of the ensuing conversation. However, it is possible to argue, as mentioned above, that "interactional authenticity" (Ellis and Barkhuizen 2005: 33) is in any respect hard to come by in learner language, considering the typical learner situation, and particularly since the learners interviewed for this corpus are all university students (see learner variables below). Since the validity of results from corpus studies may be said to be dependent on whether the language data in question reflects intended use, it is possible to argue for the validity of clinically elicited general samples as displaying language use which is similar to the type of language use a student might be asked to perform (task-driven language). It is however necessary to "acknowledge the need for multiple types of data" (Ellis and Barkhuizen 2005: 49), and the comparison of studies using multiple types of data, to further establish the effects of the situation on language output. The two corpora consist of words in context, and thus agree with Granger's (2008) demands for continuity in learner corpus data: "(…) the notion of 'continuous text' lies at the heart of corpushood" (Granger 2008: 261). The continuity of the interview as well as the few restrictions imposed on the subjects in terms of content and form, makes it justifiable to consider data from LINDSEI and LOCNEC valid data for investigating characteristics of natural conversation, and to compare findings and explanations with accounts of 'spoken conversation', such as those presented in e.g. Biber et al. (1999) (table 2.1). It is, however, important to acknowledge the

constructed nature of the data, and to be aware that, as Gilquin and De Cock points out, the nature and frequency of the language features observed, particularly those stemming from the picture description task "do not necessarily correspond to their nature and frequency in spontaneous, naturally-occurring speech" (Gilquin and De Cock 2011: 157).

The LINDSEI subcorpoa and LOCNEC display "a high degree of comparability" (Gilquin et al. 2010: 3), which is only made possible through the strict control of the corpus design. While authenticity is certainly an important consideration, comparability is pivotal for the validity of results from a contrastive study, and the LINDSEI/LOCNEC framework seems to be close to an ideal comparison, where only one variable, the native/non-native distinction, is different. With the informal interview, the LINDSEI and LOCNEC compilers wished to "ensure homogeneity in terms of text type" (ibid.), so as to allow for comparison, and at the same time impose "few constraints on language production" (ibid.), thus ensuring a high level of naturalness. Returning to Granger's task variables, the variables of LINDSEI and LOCNEC and their appropriateness in terms of comparability, authenticity and representativeness may be summarized as follows (with a few added elements not specifically considered in the previous discussion):

Task variables	LINDSEI/LOCNEC	+	-
Medium	Spoken	Typically lower level of self-monitoring than written registers; spoken corpora have not been extensively collected and analysed previously	Restricted availability of the recordings and simple transcription conventions may lead to loss of linguistic information, and can make comparison with other corpora difficult
Field	Education/Academia	Typical learner environment, familiar to the interviewees	Results may be difficult to extrapolate to different fields
Genre/text type	Informal interview	Similar to spoken conversation due to the informality, thus previous findings on spoken conversation may (cautiously) be used for comparison and explanation, restrictions makes for valid comparisons with subcorpora	Not a completely authentic sample of informal conversation; two participants only; conversational roles specified prior to conversation, the interviewer/interviewee relationship is skewed in terms of age, native language and status
Task type	Presenting a personal topic (see Appendix), informal chat prompted by the interviewer, picture description (see Appendix)	Encourages implicit performance (attention to topic rather than language), few constraints on language use, close to natural linguistic behaviour, enough restrictions for subcorpora to be comparable	Topic constraints may lead to language constraints/less authentic speech
Conditions	No reference tools available; non-surreptitious recording; each interview should last for at least 15 minutes	Absence of reference tools creates a more authentic situation, and promotes continuous language use and topic awareness rather than explicit attention to form	Non-surreptitious recording leads to increased self-monitoring

Table 3.2: *LINDSEI/LOCNEC task variables summarized*

3.2.2 Learner Variables and Representativeness as a Measure of Validity

Representativeness is important in corpus research because we generally want to report on what is *probable* in language, rather than what is *possible*. The language population a corpus aims to represent may be very diverse or very specific, and learner corpora, with the learner category as an initial narrowing of scope, usually fall somewhere between these extremes. Learner corpora thus do not aim to provide analyses applicable to all learners, but with the accumulation of various learner corpora and various studies employing these corpora, generalizations may be extended beyond the individual corpora on the basis of similarity of results.

The LINDSEI and LOCNEC corpora are, as mentioned above, compiled according to very strict and explicit design criteria. LINDSEI is divided into subcorpora according to the mother tongue background of the learners, and ideally, the variable 'mother tongue' should be the only variable separating the subcorpora, thus allowing for contrastive studies where it is possible to assume that any differences found between the subcorpora are mainly due to transfer from the mother tongue or, in the case of a LINDSEI/LOCNEC comparison, due to a discrepancy between the language proficiency of the two populations. In addition to the language itself, other aspects relating to a specific mother tongue background may also explain learner language output, such as differences in cultural norms, and these aspects should also be taken into account in an analysis of learner corpus data. Differences in educational systems and the quantity and quality of exposure to English in society are also important factors determining the learner's language output. The LINDSEI team and its contributors has opted for *external criteria*, i.e. "those which are essentially nonlinguistic" (Atkins and Clear 1992: 5), when labelling contributors to the corpus as 'advanced learners'. Following this, it was decided that all the LINDSEI interviewees should be university undergraduates in English, but a study of *internal criteria*, "those which are essentially linguistic" (ibid.), in five random samples from each subcorpora, reveals that there are clear differences in proficiency in terms of the *Common European Framework of Reference for Languages (CEF)* (Gilquin et al. 2010: 10-11). However, the samples from the Swedish subcorpus were all rated C1 or C2, thus qualifying as advanced according to both external and internal criteria, and, considering the similar social and educational conditions (learning context, see table

3.3.) in Sweden and Norway, it is possible to assume that the Norwegian samples would generate similar scores. According to the statistics, the Swedish learners distinguish themselves as being the oldest on average (27.78), having spent the highest number of months in an English-speaking country (13.78), and having had the highest number of years of English at school (9.59) (ibid.: 32-35), and these are factors that might very well have a positive effect on proficiency.

Medium of instruction	English is a compulsory subject from Grade 4 (average age: 10), but in schools English is introduced earlier. At university level, almost all specialised lectures and seminars for students of English were given in English at the time of collecting LINDSEI-SW.
Teaching focus	In primary and secondary schools, the focus is on communicative skills; receptive and productive skills are prioritised over grammatical and structural knowledge. At university more attention is given to form and accuracy, especially in written production.
Media	Television shows and films in English are subtitled, not dubbed, unless they are aimed at young children. Students are exposed to English via music, the Internet and computer games. Newspapers and books (e.g. paperback novels) in English are generally available, although this type of reading is less common among young people. A majority of the students also come in contact with English as *lingua franca* during short holiday trips.
Stays in English-speaking countries	Students at the University of Gothenburg have had the opportunity to study in Brighton for a term or more, which has been quite popular. Stays in other parts of the UK, Australia or the USA are not uncommon, whereas students at lower levels in general do not stay in English-speaking countries for long periods of time (more than one month).
Other remarks about the status of English	English has been the first foreign language at Swedish compulsory schools since the 1950s.

Table 3.3: *Societal impacts on proficiency levels for the Swedish interviewees at the time of the LINDSEI recording process (cf. Gilquin et al. 2010: 55-56)*

The LOCNEC corpus and each completed subcorpus of LINDSEI consist of 50 interviews each, which are all of similar length. The fact that the total data is divided into 50 interviews from 50 different subjects helps preventing individual

idiosyncrasies to seriously affect the corpus results. The LOCNEC interviews make up a total number of 117,417 words (learner turns only), and the Swedish subcorpus contains 71,804 words, with an average length of 1,436 words per interview (Gilquin et al. 2010: 25). The sample from the incomplete Norwegian subcorpus consists of 21 completed interviews[5] and 36,277 words of learner speech in total. These fairly small numbers will evidently affect the scope of the analysis, and may limit the validity of the results in terms of representativeness. However, considering the fact that learner corpora, and particularly spoken learner corpora, still tend to be smaller in size due to the laborious compilation process, and that, as Granger (1998b) points out, "for some linguistic studies, for instance those involving high-frequency words or structures, relatively small samples of c. 20,000 words may be sufficient" (Granger 1998b: 11), this material may still be considered valuable for the study of recurrent word-combinations.

3.2.3 Material: Summary

Following the discussions in the previous sections, it seems evident that the different subcorpora of LINDSEI cannot be extrapolated to represent 'spoken learner language' in general, but rather take part in a more complex picture of learner language behaviour. As Barlow (2005: 336) notes, results from studies of learner corpora may generally have to be regarded as preliminary until a broader range of studies have been conducted. However, it is still justifiable to assume that samples from smaller learner language populations, such as LINDSEI-SWE and LINDSEI-NOR, can tell us something about the language of advanced learners of English with a Swedish or Norwegian language background. The intricate web of task and learner variables must thus be taken into consideration at all times, and seen as comprising multiple explanations for findings, as well as possible limitations to the scope of these findings.

[5] The samples have not been subject to final corrections, thus minor transcription errors may occur in this data.

3.3 Method

3.3.1 Quantitative and Qualitative Corpus-Driven Analysis

The corpus methodology gives us the opportunity to look at larger bodies of text at the same time, and investigate the quantitative aspects of language without great difficulty. The quantitative view of texts, both in terms of general data size and quantitative searches, is thus part of what separates corpus studies from language studies that rely on other methods. However, qualitative analyses are also necessary in a corpus study, so as to avoid presenting research results as decontextualized numbers:

> "In representing grammatical differences as used in different subsections of a corpus [or different corpora], we have to make use of quantitative methods. In relating these quantitative differences to factors external to language, on the other hand, we depend on qualitative analysis" (Leech 2000: 693).

Similarly, writing on methods for detecting formulaicity in language, Read and Nation (2004) claim that "an adequate account of formulaic units as they function in language acquisition and language use can come only from a combination of quantitative and qualitative analyses" (Read and Nation 2004: 24). In addition, approaches to learner corpora may, according to e.g. Granger (1998b), be *hypothesis-based* or *hypothesis-finding*. Hypothesis-based studies build on pre-existing ideas, "generated through introspection, SLA theories, or as a result of the analysis of experimental or other non-corpus-based sources of data" (Barlow 2005: 344), while the hypothesis-finding corpus researcher "may simply decide to gather data (...) and quantify everything he or she can think of just to see what emerges" (Granger 1998b: 15). While it does not seem entirely plausible that a study can be completely free from any sort of initial hypothesis, the 'corpus-driven recurrent word-combinations method' employed in Altenberg (1998) and De Cock (1998; 2004) seems to come close to this approach. As Granger (1998b: 16) points out, "this approach is potentially very powerful since it can help us gain totally new insights into learner language" without the limitations of initial categories and assumptions. The retrieval of frequency information by the use of computer tools has, according to Barlow (2005), the advantage "that few, if any, assumptions are made about the nature of learner language" (Barlow 2005: 354). In contrast, data retrieved from a hypothesis-

based study "are viewed as valuable only in so far as they confirm or disconfirm the hypothesis" (Barlow 2005: 344). However, studies employing this approach can also run the risk of catching the "'so what?' syndrome" (Granger 1998b: 16), where frequency counts are presented without proper context, interpretation, or value beyond the numbers themselves.

Gries (2010a) distinguishes between the terms *corpus-driven* and *corpus-based* linguistics, which seems to be corresponding to hypothesis-finding and hypothesis-based methods respectively. Corpus driven studies, like hypothesis-finding studies, "aim to build theory from scratch, completely free from pre-corpus theoretical premises" and "base theories exclusively on corpus data" (Gries 2010a: 328) Gries, challenging researchers who believe corpus linguistics to be a theory rather than a methodology, argues that no study is entirely free from initial theory and assumptions, and believes that "truly corpus-driven work seems a myth at best" (Gries 2010a: 330). Hypothesis-finding studies, such as those based on the corpus-driven recurrent word-combination method, are thus, according to Gries, not truly hypothesis-finding or corpus-driven, and they probably ought not to be, if they wish to avoid purposeless "number crunching" (Aarts 2000: 7) and the closing of doors to other fields and methods. The method employed in this thesis admits to not being *purely* corpus-driven, but rather "approach corpus data with moderate corpus-external premises" (Gries 2010a: 328), an approach which Gries' would perhaps rather label as corpus-based. Similarly, its hypothesis-finding onset is unavoidably coloured by the theories and previous studies already discussed in the previous chapters. Some of the assumptions and premises underlying this study can be summarized as follows:

- The frequency, position and function of linguistic elements in use can tell us something about these elements in terms of acquisition and storage of language;
- Conclusions on the nature of formulaic language can be reached based on electronically extracted continuous word-combinations, even though many recurrent and possibly formulaic patterns may be discontinuous and/or based on abstract cognitive constructions;

- Functional and cognitive theories are the most suitable frameworks for explaining the frequency and behaviour of extracted recurrent word-combinations.

Other subjective assumptions also become apparent through all the choices made in the extracting of word-combinations and in the analysis process, as "all decisions are contingent and, in one way or another, theory sensitive" (Wray 2009: 99). This is also acknowledged by e.g. De Cock et al. (1998) in their largely corpus-driven study of 'the phrasicon of EFL learners', who acknowledge that "the filtering process required for the identification of formulaic expressions, and hence vagueness tags, is manual and to an extent subjective" (De Cock et al. 1998: 75). However, the qualitative and functional part of the analysis can also be performed with as little previous assumption as possible, as described by Erman (1987) in relation to her form-function analysis of *I mean, you know* and *you see* in English conversation:

> "Either we start with a set of functions which we wish to find realized in actual discourse, or we start at the other end, that is by making a close study of a certain linguistic item to which we then try to assign various functions. (...) The advantage of this method is that the analyst has not decided beforehand what functions to search for but carries out the analysis with an open mind" (Erman 1987: 33)

This open-mindedness towards the material is thus largely adopted as a goal for this thesis, while at the same time acknowledging the inevitable presumptions and theoretically founded decisions that are made throughout the analysis.

3.3.2 Contrastive Interlanguage Analysis

Granger (1996), following a general revival of contrastive analysis, proposed a new contrastive model for the analysis of native and learner languages, contrastive interlanguage analysis (CIA). CIA "does not establish comparisons between two different languages but between native and learner varieties of the same language" (Granger 1996: 43), which in addition to the NS-NNS comparison also includes comparisons of the non-native language of learners with different mother tongue backgrounds, such as the Norwegian and Swedish components of LINDSEI. Several researchers have embraced this approach, agreeing that "it is only by carefully comparing native-like language use and actual learner language that it is possible to identify areas in which there is still a discrepancy between the target norm and

learners' speech and areas in which learners have already approximated to the target norm" (Mukherjee 2009: 206). This comparison of native and non-native data may run the risk of becoming too normative in its approach, a pitfall often referred to as the 'comparative fallacy' (Granger 2008: 269). However, the approach also has a methodological strength, in that it makes sure that the contrast is empirically founded rather than based on intuition (ibid.). Non-contrastive studies of learner language in its own right, or contrastive studies of different learner varieties, may give us valuable insight into interlanguage behaviour, and it is important, especially in the explanatory phase of analysis, to consider learner language as a valid entity in its own right, displaying the dynamic processes common to both language acquisition and language production and development in general. It may however be argued that a contrast with some native speaker variety is needed if we are to make useful comments on aspects such as over- and underuse, transfer and avoidance, and ultimately what makes a text or feature appear 'non-native': "L1-L2 comparisons are extremely powerful heuristic techniques which help bring to light features of learner language which have not been focused on before" (Granger 2009: 18). A combination of quantitative and qualitative CIA thus seems ideal for studies of learner language, since it comprises several aspects of learner production, and points the researcher to areas where there are deviations from the native speaker norm: "both the discourse-analytical and the contrastive perspective are indispensable in analysing advanced learner language with a view to identifying areas in which the learners need to adjust or improve their performance" (Hasselgård 2009: 138).

3.3.3 Method: Summary

1. *N-gram frequency search, quantitative and preliminary qualitative CIA*
2. *Identification of possible formulaic sequences according to objective criteria and contextual information*
3. *Qualitative CIA and comparison on the basis of functional categories and usage-based theory*

Table 3.4: *Summary of method*

The initial analysis of this study is thus quantitative, inspired by the bottom-up approach of the corpus-driven 'recurrent word-combinations' method. This stage, presented in chapter 4, will help identify the recurrent word-combinations in the material, as well as quantitative deviations or similarities between the native- and non-native material. Word-combinations and frequencies presented are extracted from the corpora using the 'Clusters'-function of the AntConc 3.2.3m software, developed by Laurence Anthony (2011)[6]. In addition, examples and other contextual information are retrieved using a simple text editor, and possible explanations for the quantitative results are discussed in relation to the particular word-combinations in context. The second stage of the analysis, initiated in the analysis of chapter 4 and further developed in chapter 5, attempts to identify formulaic language on the basis of the discussion on identification of formulaic sequences in chapter 2. Finally, the last sections of chapter 5 will consider particular word-combinations of a potential formulaic nature, determining their functions and structural properties in a contextual and contrastive perspective.

[6] Available at: http://www.antlab.sci.waseda.ac.jp/antconc_index.html

4 Recurrent Word-Combinations

> "Speakers engaged in spontaneous interaction are in constant need of easily retrieved expressions to convey their intentions and reactions in discourse" (Altenberg 1998: 121)

On the basis of the theories and methodological motivations discussed in chapters 2 and 3, this chapter aims to provide a description of the occurrence of recurrent word-combinations in representative samples of native and non-native spoken English. The procedure is highly exploratory in nature, and does perhaps provide more questions than answers. The extensive set of material and categories considered also prevents a discussion and qualitative analysis of all potentially interesting quantitative findings. However, the study rather seeks to provide an overview of usage patterns of recurrent word-combinations, as well as uncovering discrepancies between the two language categories, and how these discrepancies might be explained in terms of situational and linguistic factors. Section 4.1 presents the most frequent combinations in all three corpora considered, which are further subject to a more detailed quantitative analysis in section 4.2, before a larger set of data is presented in section 4.3 along with a qualitative assessment of findings. Through these observations and analyses of the data, the aim is to discover how recurrent word-combinations function in advanced learner language specifically and compared to the native speaker norm, and how they can be said to function formulaically (or not) in the spoken texts. Chapter 5 will discuss further the formulaic aspects of particular word-combinations, and perform a more extensive qualitative study of one highly recurrent combination and its common collocations.

4.1 Corpus-Driven Frequency Search and Quantitative CIA

4.1.1 Size and Type Similarities of Highly Frequent Combinations

Simple n-gram searches of LINDSEI-SW and LOCNEC reveal that there are notable similarities between the types of highly recurrent continuous word-combinations found in the two corpora. Tables 4.1-4.3 show the twenty most frequent word-

combinations in LOCNEC and LINDSEI-SW, according to the number of words in the combination[7]. Inspired by similar searches in De Cock et al. (1998) and De Cock (2004), the tables "give us an indication of the type of result that arises from the automatic extraction of word combinations" (De Cock et al. 1998), and provide a useful starting point for a corpus-driven analysis. The rankings only serve as a partial illustration of the most frequent combinations, and since some of the combinations are equally frequent (but differently ranked), the tables do not accurately reflect the frequency data. A frequency threshold of ten occurrences was adopted for these initial searches, in accordance with the idea that a certain level of frequency is in itself a reason to consider the combinations as interesting from a phraseological point of view, as a sign of individual and collective entrenchment of a pattern (e.g. Altenberg 1998). Frequency thresholds are also useful, as noted by De Cock (2004), in order to reduce the possibility that repetitions of certain combinations are not confined to one text (interview) or one subject only. De Cock (2004) further postulates a different frequency threshold for each combination size, since "the length of recurrent word combinations is inversely related to their frequency" (De Cock 2004: 228). When listing the top twenty 2-4-word combinations in LOCNEC and LINDSEI-SW separately according to size, this consideration did not prove necessary, with the exception of 4-word combinations in LINDSEI-SW (table 4.3), where only fourteen combinations occur ten times or more. The n-gram size is set at 'greater than or equal to [number of words]' to partially illustrate the relation between frequency and size.

[7] Contracted forms, e.g. don't and it's are treated as one word in AntConc and are thus treated in the same way in the analysis. Filled pauses, eh, er, em, erm, and backchanneling, mm, uhu, mhm, are also included and treated as words.

Rank	LOCNEC	LINDSEI-SW	R.	LOCNEC	LINDSEI-SW
1	it was (757)	I think (533)	**11**	of the (310)	but I (170)
2	you know (632)	it was (419)	**12**	a lot (264)	so I (168)
3	sort of (583)	I don't (300)	**13**	but I (239)	and eh (167)
4	I mean (444)	in the (269)	**14**	don't know (237)	I mean (167)
5	I was (437)	and I (236)	**15**	a bit (229)	a lot (150)
6	I think (433)	I was (204)	**16**	so I (224)	don't know (149)
7	I don't (423)	sort of (230)	**17**	to do (220)	I don't know (139)
8	in the (416)	you know (204)	**18**	at the (217)	like that (137)
9	and I (367)	and then (184)	**19**	yeah yeah (211)	of the (132)
10	and then (345)	I I (182)	**20**	I don't know (209)	yeah I (126)

Table 4.1: *Top twenty ≥2-word combinations in NS (LOCNEC) and NNS speech (LINDSEI-SW), freq. > 10, and their raw frequencies, identical combinations underlined*

Rank	LOCNEC	LINDSEI-SW	R.	LOCNEC	LINDSEI-SW
1	I don't know (209)	I don't know (139)	**11**	you have to (56)	it was a (34)
2	a lot of (129)	a lot of (94)	**12**	don't know I (53)	when I was (31)
3	and it was (90)	I think it's (60)	**13**	I don't know I (53)	but I think (30)
4	I mean I (84)	I don't think (57)	**14**	I think I (53)	I mean I (29)
5	it was a (79)	and it was (51)	**15**	I think it's (52)	a little bit (28)
6	It was just (66)	I think it (51)	**16**	it was really (52)	all the time (28)
7	I'd like to (62)	I think I (46)	**17**	at the moment (50)	I think that's (28)
8	things like that (57)	you have to (40)	**18**	one of the (50)	and I think (27)
9	and I was (56)	I think so (38)	**19**	sort of like (50)	but it was (27)
10	I went to (56)	it was very (37)	**20**	a bit of (49)	something like that (26)

Table 4.2: *Top twenty ≥3-word combinations in NS (LOCNEC) and NNS speech (LINDSEI-SW), freq. > 10, and their raw frequencies, identical combinations underlined*

Rank	LOCNEC	LINDSEI-SW	R.	LOCNEC	LINDSEI-SW
1	I don't know I (53)	I think it was (23)	**11**	I don't know if (17)	in the middle of (10)
2	it was it was (44)	I don't know if (20)	**12**	it was really good (17)	no I don't think (10)
3	and things like that (33)	it was it was (20)	**13**	at the end of the (16)	-
4	erm I don't know (26)	or something like that (20)	**14**	or something like that (16)	-
5	at the end of (23)	and stuff like that (16)	**15**	I think I think (15)	-
6	the end of the (22)	I don't think so (16)	**16**	a lot of people (14)	-
7	a bit of a (19)	I would like to (15)	**17**	I don't I don't (14)	-
8	I think it was (19)	yeah I think so (14)	**18**	and it was really (13)	-
9	I'd like to go (19)	eh I don't know (11)	**19**	I thought it was (13)	-
10	and stuff like that (17)	I don't know what (10)	**20**	and it was like (12)	-

Table 4.3: *Top ≥4-word combinations in NS (LOCNEC) and NNS speech (LINDSEI-SW), freq. > 10, and their raw frequencies, identical combinations underlined*

In LOCNEC, there are only two recurrent 5-word combinations with a frequency of ten or more: *at the end of the,* and *you know what I mean*. A frequency threshold of five instances rather than ten retrieves two 6-word combinations from the corpus: *at the end of the day* and *teaching English as a foreign language.* Some 7-9-word combinations are also found in the native speaker corpus using this frequency threshold, e.g. *and the first thing that struck me was*. In LINDSEI-SW, there are no ≥5-word combinations with a frequency of ten or more, but a frequency threshold of 5 reveals repeated patterns like *no I don't think so, in the middle of the,* and *I can't remember the name*. Some of the even longer recurrent combinations are strictly related to the picture description task: *she shows it to her friends (and(she)), a girl sitting in a chair, (gets to) look at the picture and she*, but no 9-word combinations are recurrent in this corpus.

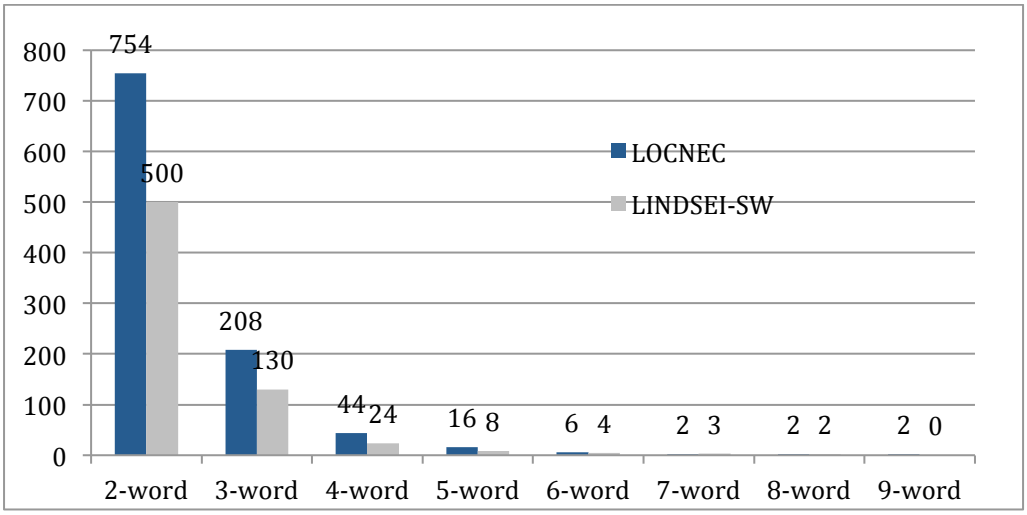


Figure 4.1: *The single most frequent 2-9-word combinations in NS speech (LOCNEC) and NNS speech (LINDSEI-SW), and their raw frequencies.*

The frequencies of the most frequent combinations show a similar pattern of distribution according to length, in both cases echoing Altenberg's (1998) observation that "continuous recurrent word-combinations in speech tend to be fairly short" (Altenberg 1998: 103). Biber et al. (1999) make a similar observation, in their investigation of 'lexical bundles' in spoken conversation, finding that there are almost ten times as many 3-word combinations as 4-word combinations in their data (Biber et al. 1999: 993). Figure 4.1 shows the raw frequencies of the most frequent type of each combination-length in the two corpora (cf. tables and descriptions above), and indicates the relationship between frequency and combination-length in LOCNEC and LINDSEI-SW. This tendency can also be seen in tables 4.1-4.3, where almost none of the twenty combinations are longer than the minimum length of the search span, with a notable exception of the highly recurrent 3-word combination *I don't know* in table 4.1, which shows a higher frequency than the majority of 2-word combinations in both corpora. The high frequency of shorter combinations is of course partly due to the fact that these may be embedded into longer combinations, such as *I think* in e.g. *I think it's, yeah I think,* and *I think so* (table 4.2), which will be further looked into in section 4.3 below. But these preliminary results also suggest that such 2-word combinations are in themselves more prevalent in the spoken repertoire of both native speakers and learners of English than are longer combinations, and include, according to Altenberg (1998: 103) "a number of phraseologically interesting idioms and collocations". The high frequency of 2-word

"fragmentary sequences" (Altenberg 1998: 102) such as *and then, so I, but I* (table 4.1) may also illustrate important organizing features of spoken discourse, which can be overlooked if only longer word-combinations are included in the search. It is thus important to also consider these combinations in studies of recurrent word-combinations, although they are often left out "because of their sheer number" (De Cock 2004: 228), which is the case in e.g. Altenberg's (1998) study.

Tables 4.1-4.3, where identical word-combinations are underlined, show striking similarities between the types of highly recurrent word-combinations in the two corpora. Searching the sample from the Norwegian LINDSEI subcorpus, we also find many of the same combinations, as presented in table 4.4[8]:

Rank	**n≥2**	**R.**	**n≥3**	**R.**	**n≥4**
1	it was (221)	**1**	a lot of (70)	**1**	and stuff like that (16)
2	I think (173)	**2**	I don't know (56)	**2**	or something like that (12)
3	and eh (157)	**3**	I went to (33)	**3**	I don't know I (11)
4	and I (150)	**4**	I think it's (31)	**4**	I think it was (10)
5	eh I (137)	**5**	it was eh (29)	**5**	-
6	in the (135)	**6**	it was a (28)	**6**	-
7	I don't (122)	**7**	and it was (23)	**7**	-
8	so I (122)	**8**	you have to (23)	**8**	-
9	I I (120)	**9**	eh it was (22)	**9**	-
10	a lot (108)	**10**	yeah yeah yeah (22)	**10**	-

Table 4.4: *Top ten ≥2-, ≥3- and ≥4-word combinations in the LINDSEI-NO sample, freq. > 10, and their raw frequencies (combinations found in the LOCNEC top 20-lists underlined)*

Many of the highly recurrent word-combinations in tables 4.1-4.4 can also be found as highly recurrent in other studies of spoken (native) English, such as *I don't know* (Altenberg 1998: 104; Biber et al. 1999: 994), *I think* (and it's expanded forms) (Altenberg 1998: 113; Biber et al. 1999: 1002) and *or something like that* (Altenberg 1998: 117; Biber et al. 1999: 1012), which suggests that these combinations are somehow characteristic of spoken English conversation, and also that they are not

[8] Considering the small size of the Norwegian sample, only the top ten combinations were included in the search.

strictly a product of the LINDSEI/LOCNEC contextual factors. *I think* is also reported to be the single most frequent *I* + verb combination in the spoken components of the Corpus of Contemporary American English (COCA) (Davies 2008) and the British National Corpus (BNC), whereas *I don't know* is the single most frequent negative collocation in these corpora (Baumgarten and House 2010: 1186). Altenberg's material, the London-Lund Corpus of Spoken English, consists of spontaneous conversations, most of them recorded surreptitiously, and thus represent more naturally occurring speech than LINDSEI/LOCNEC. Biber et al.'s (1999) corpus of spoken conversation is recorded in a similar fashion, and their ensuing analysis shows that "by far the most prevalent type of lexical bundle in conversation is a clause fragment, consisting of a subject pronoun followed by a verb phrase", and that "in many cases, the verb phrase is extended by the beginning of a following complement clause" (Biber et al. 1999: 1002). Similarly, Altenberg (1998), adopting a linear distribution scheme to describe these multiple clause constituents, find many subject-verb combinations in clause initial and medial position, forming "the springboard of utterances leading up to the communicatively most important - and lexically most variable - element" (Altenberg 1998: 113). This tendency can also be seen across all three language populations above, and will be further discussed in the following sections.

4.1.1.1 Unfilled pauses, filled pauses and repetitions

One precaution must be made in relation to the length and general build-up of word-combinations in this material, and that is the presence of unfilled pauses, signalled by a punctuation mark in the LINDSEI and LOCNEC transcriptions. These pauses are retained in the texts, and regarded as interruptive markers by AntConc in the n-gram search. Thus tokens of combinations which are on the lists of the most recurrent word-combinations (examples 1a and 2a), are not included in the frequency lists if interrupted by one or more unfilled pauses (examples 1b and 2b):

(1) *a.) I don't know I don't know it depends you know it depends how good it is you know* (LOCNEC)

b.) I don't know . I really don't know (LOCNEC)[9]

(2) a*.) so it was it was quite relevant to us* (LOCNEC)

b.) it was yeah it was quite high up it was .. it was beautiful (LOCNEC)

Although reports on the presence or absence of pauses as a defining feature of formulaicity are inconclusive, "confidence in the validity of pause location as an indicator of a boundary has increased to the point where pauses are now being used to help identify formulaic sequences" (Wray 2009: 104), and it thus seems reasonable to consider them as natural interruptions to sequences in n-gram frequency searches. This issue is also addressed by De Cock et al. (1998) in relation to the inclusion of *filled* pauses, *erm, em* etc., in the electronic retrieval of recurrent word-combinations, who conclude that "the prefabricated nature of formulae and recurrent word combinations in general makes it (...) valid to assume that their production may not be interrupted by extraneous hesitation features" (De Cock et al. 1998: 71). Keeping them in the search, however, has implications for the retrieved results, and it is important to acknowledge that interesting patterns and tokens of recurrent word-combinations may be left out as a consequence. This is also the case with discontinuous and variable sequences, which will not be considered here unless their different types appear in the frequency searches.

Repetitions*, I I* (LINDSEI-SW&NO), and filled pauses, *erm I don't know* (LOCNEC), appear in some of the highly recurrent word-combinations in the corpora, and they are likely to be present also in the recurrent combinations which fall outside of the top 20 lists. The recurrence of filled pauses in LINDSEI and LOCNEC must perhaps be considered with particular caution, as they are transcribed in four different ways in the

[9] Examples from LOCNEC and LINDSEI are marked with the tags <A></A> (interviewer) and <B></B> (interviewee) in cases where both participants are represented, and left without tags if the example includes text from the interviewee only. Most examples do not adhere to any pre-defined clausal or conversational boundaries, but are cut off after the necessary contextual information is judged to be provided. The tag <overlap /> is removed from examples where the overlapping speech is not included or where the overlap is not considered significant for the subject matter of the example. See Appendix for detailed transcription conventions used in the transcription of the LINDSEI corpora.

corpora, *eh, er, em* or *erm,* according to length and sound profile. For instance, the combination *and + [filled pause]* in the Swedish LINDSEI can be found with all four filled pauses represented, and thus the combination *and eh,* which occurs in the top 20 list in table 4.1 could be considered as an even more frequent combination if the other filled pauses were taken into account. Following this, other combinations containing filled pauses might have been higher up on the top 20 lists, had these four items in some way been combined. Filled pauses are included in the frequency search in order to illuminate some of the functional properties these items may possess in combination with other words and word-combinations, such as indicating encoding problems at clause beginnings (De Cock 2004: 233) and "furthering smooth and effortless conversation" (Kjellmer 2003: 171). Trying to include filled and unfilled pauses as a stable component of formulaic language would undoubtedly cause difficulties, but since they are such a prevalent part of spoken language, and since they do seem to follow certain patterns relevant to formulaicity, it seems unwise to disregard them. This is also partly true for repetitions, and it is, in practical terms, difficult to exclude these altogether from the frequency lists, since some seemingly non-repetitive combinations such as *I don't know I* (LOCNEC, LINDSEI-NO) may occur as a frequent combination partly as a result of repetition of a word-combination (see example 1a).

Although 'dysfluencies' or 'hesitation items' such as repetitions and pauses may be considered to be "of little phraseological interest" (Altenberg 1998: 103), these items may, as mentioned, also be of interest for a better understanding the overall organization of spoken discourse, and highlight problems of planning pressure in both native and learner language. Kjellmer (2003) finds in his study of filled pauses in spoken material from the Cobuild corpus of native speaker English, that the word most frequently occurring before the filled pauses *er* and *erm* is the conjunction *and*, and that the personal pronoun *I* is the most frequent collocate for the position following immediately following them. In tables 4.1-4.3, we can see that this tendency is apparent in LOCNEC and both the learner corpora, with *and eh* (LINDSEI-SW&NO), *eh I* (LINDSEI-NO) and *erm/eh I don't know* (LOCNEC, LINDSEI-SW) among the most frequent combinations containing filled pauses, in addition to *it was eh* and *eh it was* in the LINDSEI-NO sample (table 4.4). Kjellmer

deduces from his findings that "one main function of er(m) thus seems to be to introduce what I will loosely call a new 'thought unit', a word, a phrase and sometimes a whole clause" (Kjellmer 2003: 174). Following up on Kjellmer's findings, Tottie (2010) suggests the term 'planners' to refer to filled pauses, rather than the more widespread, and more negatively charged term 'hesitation items'. Even though these planners seem to serve important functions in facilitating conversation, Kjellmer also concludes that "since we are most of the time unaware of the [filled pauses], their (moderate) use will not normally affect adversely our impression of a speaker's fluency or eloquence" (ibid.: 191). It is possible to assume that this is true also for repetitions, and it is also likely that both of these language features, when overused or 'misused' relative to a native speaker norm, will make the listeners become aware of the dysfluencies, and that this will have an impact on our impression of a *learner's* fluency or eloquence. This potential source of non-nativeness is referred to in e.g. Brand and Götz's (2011) pilot study on LOCNEC and the German component of LINDSEI, which did not reach any firm conclusions. Considering the planning function which might be relevant for these items, it is interesting to compare the findings from the Swedish and Norwegian learner corpora above with the top 20 3-word combinations from the French component of LINDSEI, as presented in De Cock (2004):

	LINDSEI-FR		
Rank		**R.**	
1	I don't know	**11**	and er we
2	I I I	**12**	and so on
3	and it was	**13**	no no no
4	and er well	**14**	but I I
5	the the the	**15**	to to to
6	and er I	**16**	I I was
7	and er the	**17**	yes yes yes
8	it was really	**18**	a lot of
9	it was er	**19**	I would say
10	it was a	**20**	I went to

Table 4.5: *Top 20 3-word combinations in LINDSEI-FR (cf. De Cock 2004: 228)*

Compared to the Swedish and Norwegian learners, and to the native speaker corpus, table 4.5 shows a greater use of combinations containing repetitions and filled pauses among the French learners. In total, according to De Cock (2004), the French learners

in LINDSEI "use approximately 3 to 4 times as many sequences that contain repeats and/or hesitation items as native speakers" (De Cock 2004: 233). The preliminary results from the Swedish and Norwegian data suggest that the number of repetitions and filled pauses is lower in these corpora, which is an interesting base for further research. It is possible to hypothesize that if the usage patterns of pauses and repetitions in LINDSEI-SW and LINDSEI-NO are more similar to those found in native English speech, this is a reflection of a higher level of general proficiency among these learners, where planning of a new 'thought unit' and retrieval of the desired lexical form, whether that be a word, a phrase, a formulaic sequence or a clause, comes easier. Considering the common conception about many formulaic sequences, similar to that of pauses and to a certain extent repetitions, that they free up cognitive space for the production of more complex ideas or linguistic patterns, it is also possible that the Norwegian and Swedish learners, more advanced and exposed to word-combinations in English, make a greater use of some formulaic sequences for these planning purposes, rather than pauses and repetitions, and thus show a greater fluency. Raupach (1984), in his case study of the language development of German learners of French, suggests that such a change in planning behaviour does occur with increased proficiency:

> "Part of the planning activities that previously had been reserved for silent and filled pauses is now processed in connection with other hesitation phenomena and at other places than before. This shift in the placement of 'islands of reliability' may be the acquisition of new organizers leading to a preferred set of formulaic schemata (c'est; il y a; en ce qui concerne) and collocations" (Raupach 1984: 135).

The further investigation will thus not concern itself with pauses and repetitions in isolation, but will consider them as part and context of candidates for "recurrent speech patterns activated by an individual speaker to overcome or to avoid difficulties in organizing his performance" (Raupach 1984: 134).

4.1.2 Preliminary N-Gram Search: Summary

The above observations and the results from tables 4.1-4.4 may imply that the Swedish and Norwegian learners are so advanced, relative to a native speaker norm, that their use of recurrent word-combinations, and in turn their language production in general, is highly similar to the native speakers' in this otherwise controlled setting. However, the top 20- and top 10-lists do not tell us enough about the contrasts

between the usage patterns in the corpora, in terms of relative frequencies (quantity) and functional properties (quality), to justify such a claim. Leaving the results at this stage might conceal important differences in the data, and the further analyses below will shed light on some of the quantitative and qualitative differences between these and other word-combinations in the data. However, it seems reasonable to assume, on account of these preliminary results, that the Swedish and Norwegian learners are very familiar with spoken English, in such a way as to adopt, memorize and make use of some of the very common word-combinations in English conversation. The Swedish learners are also independently rated as the most advanced learners out of all the mother tongue backgrounds collected in the LINDSEI project, and it is likely that the Norwegian learner data would generate similar scores, as mentioned in section 3.2.2. The assumption of a relationship between proficiency and types of highly recurrent word-combinations may however also be flawed, since some of these combinations, on the basis of their high frequency, might very well be likely candidates for memorization and production at a very early stage of the learning process, as pointed out by Schmitt and Carter (2004): "In L2 acquisition, formulaic sequences are also relied on initially as a quick means to be communicative, albeit in a limited way" (Schmitt and Carter 2004: 11). Such 'phrasal teddy bears' might thus be prevalent in the same way as lexical ones: "learnt early, widely usable, and above all safe (because they do not show up as errors)" (Hasselgren 1994: 250). More comprehensive contrastive analyses of these top 20 lists from the different LI subcorpora of LINDSEI would, considering the differences in proficiency displayed in these corpora, be revealing in this respect.

The quantity and quality of the top twenty frequency lists support the phraseological notion that "words belong with other words not as an afterthought but at the most fundamental level" (Wray 2002: 13) - both in native speaker speech and the speech of advanced learners. They indicate that native and learner language do not solely consist of 'individual building blocks' assembled according to predefined rules and semantic information, but rather appear to be produced partly on the basis of larger, previously encountered and memorized sequences. It is possible to argue that it is the very specific situational context which prompts the use of specific words in combination in the case of LINDSEI and LOCNEC, particularly considering the

combinations produced during the picture description task. It may also be argued that e.g. the perceived sequence of concepts which prompts the combination *she shows it to her friends* (as illustrated in the cartoon, see Appendix), is almost impossible to render in any other way, considering the semantic contents of the individual words, and that the production of this string by three of the Swedish learners is due to an identical organization of concepts, both spatially and temporally, rather than a language specific memorization of word-combinations. This consideration, however, becomes less of a problem when considering many of the other recurrent word-combinations in the corpora, e.g. *I don't know* and *I think*, which are not as strongly tied to and prompted by the situational context, and which are also intuitively easier to replace with other combinations to achieve the same propositional or functional content. These combinations are also noticeably frequent, even at the stage of only dealing with raw frequencies, which suggests that they appear in several contexts and perform a wider range of functions than those related to the picture description task (see section 4.3.4 below for a further discussion on the picture description task). They might also, by virtue of their semantic or pragmatic unity, be considered to be better examples of formulaic sequences, according to the working definition postulated in section 2.2.2 above:

> A formulaic sequence is defined as the co-occurrence of two or more consecutive lexical items which functions as one semantic/pragmatic unit in a clause or sentence, which is, or appears to be, prefabricated and conventionalised, and whose frequency of occurrence is larger than expected on the basis of chance.

The preliminary results thus lead to a number of questions, which may be explored further through a more in-depth analysis of the data. Some of these questions may be summarized as follows:

- Why are there so many similarities between the most frequent 2-4-word combinations in LOCNEC, LINDSEI-SW and the LINDSEI-NO sample?
- Are there also noticeable *differences* in the relative frequencies of the highly recurrent word-combinations occurring in the corpora, and if so, what does this entail?
- Are the most frequent word-combinations types typically embedded into other frequent combinations?

- Are there noticeable differences in the functional patterns (usage) of the highly recurrent word-combinations in the corpora, and if so, is this a result of proficiency levels, or other factors?
- Can the highly recurrent word-combinations found in the corpora justifiably be labelled as formulaic sequences, according to the working definition of this thesis?

The next section will examine further the quantitative aspects of these questions.

4.2 Highly Frequent Word-Combinations: Further Frequency Findings

In order to level out the differences in sample size between the NS and NNS corpora, relative frequencies have to be calculated and normalized. Many comparative studies of corpus data are calculating relative frequencies based on the total number of words in the corpus samples which are to be compared. This approach has been criticised from a statistics point of view, particularly when employed in corpus studies of grammatical forms, and it is suggested that relative frequencies of e.g. the occurrence of the present perfect in a corpus should be expressed on the basis of the number of verbs rather than words in the data (Gries 2007: 112), to make results more valid for interpretation and comparison. Ball (1994) also questions word count as a valid frequency metric:

> "In a word-based frequency analysis, to say that a phenomenon occurs with equal frequency in two samples is to say that equal amounts of text, measured in words, will yield the same number of tokens. But relative frequency should be a measure of the number of times something occurs within the number of opportunities for it to occur" (Ball 1994: 297).

In a predominantly bottom-up study of recurrent word-strings, however, the total word-count seems to be the most appropriate reference for calculating relative frequencies, since these structurally variable strings extracted cannot (and should not) be grouped into any other corpus-external category, and since it seems impossible to predict the number of opportunities for a non-specified recurrent word-combination to occur. Alternatively, it is possible to calculate the number of times a given n-gram occurs out of the total possible recurrent n-grams, e.g. the 2-gram *I think* divided by the total number of recurrent 2-grams in the corpus. This approach would however partly exclude the recurrent/non-recurrent factor of the analysis, and rather consider

which lexical combinations are the most frequent compared to other combinations, rather than compared to the corpus as a whole. In addition, repetition of linguistic items may be said to potentially occur at any position in language production. Relative frequencies and statistical tests based on the number of times word combinations occur in relation to the total word-count are thus considered to be informative for the purpose of this study, albeit with the precaution that the results of statistical tests for this kind of frequency data may be less reliable and transferable than for other linguistic phenomena. This approach is also partly in line with other contrastive studies of recurrent word-combinations, such as Biber et al. (1999), De Cock et al. (1998), Granger (1998a) and Dahlmann and Adolphs (2009).

4.2.1 Frequency Distributions

Tables 4.1-4.4 showed many similarities between the types of highly recurrent continuous word-combinations in LOCNEC, LINDSEI-SW and the LINDSEI-NO sample, but the raw figures did not allow for a valid comparison of their frequencies, considering the differences in corpus size of about 45,600 words between LINDSEI-SW and LOCNEC, and 81,100 words between the LINDSEI-NO sample and LOCNEC. The identical 2-5 word-combination types found in LINDSEI-SW and LOCNEC are presented in table 4.6, together with their relative normalized frequencies per 10,000 words[10], and ranked according to their frequencies in the native speaker corpus. Table 4.7 shows the corresponding results for LINDSEI-NO and LOCNEC. The remaining word-combinations in the LINDSEI-SW top 20-lists which are not identical to any of the top twenty combinations in LOCNEC are still similar to the remaining LOCNEC combinations in form, and these combinations will be further discussed in relation to the extended presentation of the word-combinations in the corpora in section 4.3 below.

[10] This normalization figure seems appropriate considering the small total number of words in the corpora.

WORD-COMBINATION	LOCNEC		LINDSEI-SW		
	n	***n* per 10,000**	***n***	***n* per 10,000**	***n* per 10,000 - *n* per 10,000**
2-GRAMS					
it was	757	64.5	419	58.4	**6.1**
you know	632	53.8	204	28.4	**25.4**
sort of	583	49.7	230	32.0	**17.6**
I mean	444	37.8	167	23.3	**14.6**
I was	437	37.2	204	28.4	**8.8**
I think	433	36.9	533	74.2	**-37.4**
I don't	423	36.0	300	41.8	**-5.8**
in the	416	35.4	269	37.5	-2.0
and I	367	31.3	236	32.9	-1.6
and then	345	29.4	184	25.6	3.8
of the	310	26.4	132	18.4	**8.0**
a lot	264	22.5	150	20.9	1.6
but I	239	20.4	170	23.7	-3.3
don't know	237	20.2	149	20.8	-0.6
so I	224	19.1	168	23.4	-4.3
3-GRAMS					
I don't know	209	17.8	139	19.4	-1.6
a lot of	129	11.0	94	13.1	**-2.1**
and it was	90	7.7	51	7.1	0.6
I mean I	84	7.2	29	4.0	**3.1**
it was a	79	6.7	34	4.7	**2.0**
you have to	56	4.8	40	5.6	-0.8
I think I	53	4.5	46	6.4	-1.9
I think it's	52	4.4	60	8.4	**-3.9**
4-GRAMS					
it was it was	44	3.7	20	2.8	**1.0**
I think it was	19	1.6	23	3.2	**-1.6**
and stuff like that	17	1.4	16	2.2	-0.8
or something like that	16	1.4	20	2.8	**-1.4**

Table 4.6: *Highly recurrent word-combinations (freq. >10) occurring in both LOCNEC and LINDSEI-SW, raw frequencies, normalized frequencies (per 10,000 words) and their difference (differences >5/2/1 marked in bold)*

WORD-COMBINATION	LOCNEC		LINDSEI-NO		
2-GRAMS	***n***	***n* per 10,000**	***n***	***n* per 10,000**	***n* per 10,000 - *n* per 10,000**
it was	757	64.5	221	60.9	3.6
I think	433	36.9	173	47.7	**-10.8**
I don't	423	36	122	33.6	2.4
in the	416	35.4	135	37.2	-1.8
and I	367	31.3	150	41.3	**-10.0**
a lot	264	22.5	108	29.8	**-7.3**
so I	224	19.1	122	33.6	**-14.5**
3-GRAMS					
I don't know	209	17.8	56	15.4	**2.4**
a lot of	129	11	70	19.3	**-8.3**
and it was	90	7.7	23	6.3	1.4
it was a	79	6.7	28	7.7	-1.0
I went to	56	4.8	33	9.1	**-4.3**
you have to	56	4.8	23	6.3	-1.5
I think it's	52	4.4	31	8.5	**-4.1**
I think it was	19	1.6	10	2.8	-1.2
4-GRAMS					
and stuff like that	17	1.4	16	4.4	**-3.0**
or something like that	16	1.4	12	3.3	**-1.9**

Table 4.7: *Highly recurrent word-combinations (freq. >10) occurring in both LOCNEC and the LINDSEI-NO sample, raw frequencies, normalized frequencies (per 10,000 words) and their difference (differences >5/2/1 marked in bold)*

This conflation of tables 4.1-4.4 into two confirms yet again that the NS and NNS corpora have strikingly many highly frequent combinations in common, particularly concerning 2-word-combinations, where almost all the top 20 and top 10 combinations are identical to those found in the native speaker corpus. The difference based on normalized frequencies in the right-most column show that there are also similarities in the frequency of distribution of these word-combinations, albeit with some noticeable exceptions. The Swedish non-native speakers seem to underuse some highly frequent combinations compared to the native speakers, particularly *it was, you know, sort of, I mean, I was* and *of the* (with a positive difference of >5). The Norwegian learner data does not show a similar pattern of underuse, but there are conflating patterns of overuse, particularly concerning the highly frequent *I think* (with a negative difference of >10 in both corpora). By applying the chi-square test to the combinations with the highest differences, we can find out whether the observed potential overuse and underuse of word-combinations are significant findings also in statistical terms. Chi-square tests can compare observed frequencies in two or more sets of data, and calculate how statistically significant any differences between these

figures are (cf. e.g. Meyer (2002:122-132). A high chi-square value may thus tell us that the observed differences are not solely due to chance, but rather appear as a result of the different properties of the two data sets, in this case (predominantly) the native/non-native distinction. As mentioned above, recurrent word-combinations fall between categories in terms of statistical information, and this is also apparent in the chi-square test, where word-combinations must be treated as one word and calculated on the basis of the total number of words. However, this approach is similar for the two corpora, ensuring some validity in this respect.[11]

[11] The chi-square results are manually calculated in Microsoft Excel, according to instructions provided by e.g. Anatol Stefanowitsch. (2004): http://www-user.uni-bremen.de/~anatol/qnt/qnt_dist.html.

WORD-COMBINATION	LOCNEC	LINDSEI-SW		LINDSEI-NO	
	***n* per 10,000**	***n* per 10,000**	**X^2**	***n* per 10,000**	**X^2**
Potential underuse					
it was	64.5	58.4	2.70	60.9	0.55
you know	53.8	28.4	**65.43**	16.5	**85.96**
sort of	49.7	32.0	**32.34**	14.9	**81.16**
I mean	37.8	23.3	**29.33**	2.5	**117.73**
I was	37.2	28.4	**10.24**	28.1	**6.57**
I don't	36.0	-	-	33.6	0.45
of the	26.4	18.4	**3.88**	21.2	2.96
a lot	22.5	20.9	0.52	-	-
I don't know	17.8	-[12]	-	15.4	0.9
I mean I	7.2	4.0	**7.24**	/ [13]	/
it was a	6.7	4.7	2.97	7.7	0.39
I went to	4.8	1.7	**4.37**	-	-
it was it was	3.7	2.8	1.22	/	/
Potential overuse					
I think	36.9	74.2	**122.40**	47.7	**8.25**
I don't	36.0	41.8	**3.88**	-	-
and I	31.3	32.9	0.36	41.3	**8.42**
a lot	22.5	-	-	29.8	**6.09**
so I	19.1	23.4	**4.02**	33.6	**26.13**
I don't know	17.8	19.4	0.59	-	-
I went to	4.8	-	-	9.1	**43.2**
I think it's	4.4	8.4	**10.24**	8.5	**8.70**
I think it was	1.6	3.2	**5.04**	2.8	1.9
or something like that	1.4	2.8	**4.74**	3.3	**5.76**
and stuff like that	1.4	2.2	1.56	4.4	**11.33**

Table 4.8: *High-difference combinations and chi-square results (cf. tables 4.6 and 4.7), significant values ($p<0.05$, d.f.=1) marked in bold*

Since there did not seem to be evidence of underuse of the most frequent patterns in LINDSEI-NO compared to the native speaker corpus, I wished to see whether the word-combinations that were used less in the Swedish subcorpus could be found further down on the frequency list in the Norwegian sample, and thus providing firmer evidence for underuse of these particular patterns among learners in general. The high-difference combinations indicating overuse in LINDSEI-NO, and which could not be found in the Swedish top 20-lists, were also added to the list for comparative reasons. Table 4.8 shows the high-difference combinations in LINDSEI-SW and LINDSEI-NO and their individual chi-square results as compared with

[12] - = discrepancy between overuse/underuse in the LINDSEI corpora.
[13] / = not recurrent in the corpus.

LOCNEC. The combinations *I don't know, a lot, I went to* are found in both the 'underuse' and 'overuse'-columns, as they are differently distributed in the two LINDSEI-corpora.

The chi-square scores show that most of the combinations of different sizes that were singled out on the basis of a high frequency difference (>5, >2 and >1, respectively), are also significantly different according to the statistics, with the exception of *it was a, it was, it was it was* which fall below the 5 % error probability threshold (3.841[14]) in both the learner corpora. The results of underuse for *you know, sort of* and *I mean* are all *highly* significant in both corpora, with an error probability rate of 0.1% (10.828). The bottom column shows that the 2-word combination *I think* is highly overused in the Swedish corpus, and that the two 3-word combinations containing *I think* are also significant. *I think* is also significantly overused by the Norwegian learners, who further seem to highly overuse the combinations *and stuff like that, I went to*, and *so I.*

An interesting find when comparing the LINDSEI-corpora is that even though the Swedish learners do underuse the combinations *you know, sort of* and *I mean*, the Norwegian learners seem to use these combinations much less frequently. The related combination *I mean I*, which is also underused in the Swedish corpus, is not found at all in the Norwegian sample. These differences might be due to individual learner preferences, as the low number of interviews in the corpora, particularly in the Norwegian sample, makes the results vulnerable to such factors, but they might also be due to differences in proficiency between the Swedish and Norwegian learners, or to transfer from an existing inventory of similar expressions and their use in Swedish. Taking individual preferences into account by calculating the means of occurrences in each interview would help level out this particular factor, and a more in-depth analysis of the differences between the Swedish and the completed Norwegian corpora would help answer the questions of quantitative differences. The further analyses of this study will however focus mainly on the native/non-native distinction,

[14] d.f.=1.

i.e. on similarities found in LINDSEI-SW and -NO, and, in turn, their similarities and differences as compared to LOCNEC.

4.2.2 Summary and Preliminary Conclusions

Only the most frequent recurrent word-combinations of different sizes have been considered so far, and particularly the combinations that occur often in both the native and the non-native speaker corpora. The combinations with a high chi-square test score in table 4.8 stand out as being overused or underused compared to the native speaker corpus, even though they are all very frequent within all three corpora. The question asked in section 4.1.2 on relative frequencies is thus answered, and tables 4.6-4.7 show that some of the highly frequent combinations also have very similar distribution patterns to LOCNEC, such as *a lot* (LINDSEI-SW)*, you have to* and *I don't know* (both NNS corpora). By virtue of being overused or underused, however, the highly frequent word-combinations with statistically significant chi-square resuls in table 4.8, e.g. *I think, I mean* and *or something like that*, appear interesting both from a phraseological point of view, as potential formulaic sequences common to English speech, and from the prospect of second language acquisition, as potential contributors to perceived aspects of non-nativeness in advanced learner English speech.

Having established that there are differences in the frequency distribution of highly frequent combinations, the question remains as to what these differences entail. In his study of vocabulary in the ICLE corpus of written learner English, Ringbom (1998) finds that high-frequency verbs such as *think* are found more often in the Swedish subcorpus (and all the other subcorpora) than in the native speaker reference corpus LOCNESS, and suggests that part of the explanation for this overuse may be their taking part in recurrent word-combinations like *I think* (Ringbom 1998: 44). Ringbom believes that patterns of overuse of a limited number of words and word-combinations indicates an "insufficient and imprecise (...) use of the resources available in English" (ibid.: 51), making written learner language appear "vague and stereotyped" (ibid.: 49). It is possible that the overuse of combinations like *I think* found in both written and spoken language is due to an overgeneralization of these combinations, which in turn appear as a result of a restricted lexical inventory. The fact that the word-

combinations that are shown to be overused in table 4.8 are also highly frequent in the native speaker corpus makes it possible that this overuse might not be consciously picked up as non-nativelike by listeners, but their functional profile as well as the possibility of a corresponding underuse of other word-combinations, might however contribute to this non-nativelike impression. The linguistic context of these overused word-combinations and possible explanations for their underuse is discussed in the following section.

Ringbom's notion of 'vague' learner writing is perhaps a different sort of 'vagueness' than the one commonly explored in relation to spoken language, where vague language seems to be a valued trait: "vagueness or lack of precision is one of the most important characteristics of informal interaction" (De Cock 2004: 236). De Cock claims, in relation to the French learners' underuse and misuse of 'vagueness tags' like *and so on, (or) something like that* and *sort of* in informal speech, that "learners' preferred sequences are (...) less interactional and involved in nature than native speakers'" (ibid.: 235). The importance of vagueness markers in conversation as opposed to written registers is also pointed out by Biber et al., in relation to one of their external determinants of conversation, 'conversation avoids elaboration or specification of meaning':

> "Seen from the vantage point of written language, with its emphasis on specificity, such vagueness appears to be a culpable lack of precision. But, from the viewpoint of conversational partners, greater precision would not only be superfluous, but it would also need more processing and delay the ongoing dynamic of the conversation. Being inexact about values and opinions, like being unspecific in reference, is a strategy which relies on an implied sharing of knowledge and experience" (Biber et al. 1999: 1045).

The Swedish and Norwegian learners seem, on the basis of the results presented above, to underuse some frequent markers of vagueness, like *sort of,* but also slightly overuse others, like *and stuff like that* (significant overuse only in LINDSEI-NO) and *or something like that* (which is not found at all in the French subcorpus, cf. De Cock (2004)). It is possible that frequently occurring word-combinations like *I think* and *I don't know*, although unusual when appearing often in written registers, could also be used and interpreted as markers of vagueness in speech, and employed by learners as a conscious or unconscious compensation strategy for lack of other ways of being 'inexact about values and opinions', as seen in (3) and (4):

(3) *and even though we were just going to: a restaurant or: to a shop . we have to dr= take the car we couldn't walk anywhere . we I (eh) noticed a sign . (eh) it was a big sign . beware pedestrians <begin laughter> could come here you know so it was <end laughter> it was quite strange. I think* (LINDSEI-NO)

(4) *no it's (eh) (eh) I think there's . too little green areas and too (em) . (erm) yeah* (LINDSEI-SW)

If these combinations are easy to retrieve from the mental lexicon in real-time processing situations due to their entrenchment as formulaic patterns, it is likely that they appear often, and perhaps in contexts where native speakers would make use of other combinations from their larger lexical and formulaic repertoire. Their prevalence in written language may thus be due to a lack of register awareness, or to the fact that they are so easily retrieved that they overshadow any alternative word-combinations the learner might have used instead. Since these are high-frequency word-combinations also in native speaker speech, the extended use might be due to a 'chunking process' in both the native and learner mind, partly as an effect of the high frequency:

> "Some of the effects of chunking are rather subtle: small phonetic adjustments, most of which are variable; possible slight increases in accessing speed; and recognition by speakers that certain combinations are conventional. However, with increasing frequency, other more dramatic changes occur in chunks. These include changes in morphosyntactic structure, *shifts in pragmatic nuances* and functions and change in semantics" (Bybee 2010: 44; my italics).

To answer the remaining questions posed at the end of section 4.1, on the embedding, functional patterns and formulaic status of frequent word-combinations, an extended data set must be analysed. Considering a larger stock of word-combinations will provide further information on variation and the collocational frameworks of the highly frequent combinations, and thus show whether or not they owe their frequency partly due to embedding into longer and frequent combinations. Taking less frequent combinations into account will also test the claim that recurrent word-combinations in the NNS material are highly similar to those extracted from the NS corpus, and take into account the combinations that do not correspond in the material and which have not been considered thus far.

4.3 Extending the Material

Below is an overview of recurrent 2, 3 and 4-word-combinations in the LOCNEC and LINDSEI-SW corpus, together with the recurrent combinations in which they are embedded. The frequency thresholds were set at 77 and 50, 23 and 15, 7 and 5 respectively, in order to limit the material to a manageable size, to partly level out the differences in size between the two corpora, and considering the general size/frequency discrepancy. The search retrieved 120 2-gram types, 84 3-gram types and 97 4-gram types from the LOCNEC corpus, and 114 2-gram types, 85 3-gram types and 113 4-gram types from the Swedish LINDSEI. Since the Norwegian sample is incomplete and small in size, it was not included in this more comprehensive overview, but examples from this material will be further employed in the contextual analysis below, and in the discussion of formulaicity in chapter 5.

The overview is inspired by Altenberg's (1998) classification of combinations according to structural form and pragmatic and textual function, and is based on the formal and clausal properties of the combinations. This is also a similar, though simplified, approach to the one employed by Biber et al. (1999) in their overview of lexical bundles in conversation (Biber et al. 1999:1001-1014). Combinations that are, at surface level, structurally complete as full clauses and those including more than one clause constituent are presented in tables 4.9 and 4.10, while incomplete clauses and phrases are presented in tables 4.15 and 4.16, along with complete single clause constituents. The combinations have also been ordered according to the presence of pronouns, classifying incomplete combinations like *it's a* together with other combinations which also include the third person pronoun in tables 4.9 and 4.10, rather than in the incomplete clauses-tables. The combinations strictly related to the picture description task are listed separately (tables 4.17-4.18), as are clear cases of repetition and the combinations with filled pauses (tables 4.19-4.20).

A purely form-based classification is chosen at this point rather than a combined form-function classification, like the one presented in Altenberg (1998). Altenberg sorts his material into general functional types based on interaction, such as responses (*yes of course*), comment clauses (*I should think*) and vagueness tags (*or something like that*), as well as discourse organizational functions like frames (*and you know*) or

stems (*I think/thought (that))* (ibid.). Through this classification word-combinations are ascribed to functional categories of a textual or interactive nature. A similar sorting of the combinations in the material presented below appears to be a difficult and somewhat uninformative undertaking, particularly considering the fact that 2-word combinations are also included in this study. Some word-combinations, particularly among the incomplete clauses and phrases (tables 4.15 and 4.16), are difficult to classify according to function, as they do not appear to function as a unit, semantically or pragmatically. Other combination tokens may perform multiple functions simultaneously, or different functions in different contexts, such as e.g. *I don't know*, which "does not have a single function but is characterised by its broad spectrum of uses" (Aijmer 2009: 156).

(5) *oh actually it's . I don't know if it counts as a minor itself it's part of English literature . erm* (LOCNEC)

(6) *erm I don't know I might . I might take a year out in France* (LOCNEC)

In (5), *I don't know*, which is embedded in the recurrent combination *I don't know if*, functions as an epistemic stem (Altenberg 1998: 114), softening the assertion of the following subclause. In (6), however, the combination, together with the filled and unfilled pause and the repeated combination *I might*, seems to take part in a textual frame (ibid.: 112) which shows weaker epistemic content, and rather qualify as what Aijmer (2004) would label a 'pragmatic marker', having "the function of checking that the participants are on the same wavelength, or of creating a space for planning what to say, making revisions, etc." (Aijmer 2004: 177). Wray (2002) discusses the varying functional aspects of formulaic language in relation to the definition and identification of formulaicity, claiming that "for at least some formulaic sequences associated with social functions, the word-string itself is only part of the interaction, the remainder being achieved by other aspects of behaviour and by appropriacy of context" (Wray 2002: 55), a premise which makes it difficult to perform a general functional classification purely on the basis of the form of the word-combinations. In addition, it seems particularly important to consider more contextual information prior to a functional classification in the case of the learner material, where the functions of recurrent word-combinations may vary greatly from the nativelike use seen in e.g. Altenberg's material and in LOCNEC. Taking linguistic and propositional context

into account, a functional approach based on Altenberg's categories and the categorisation of other comparable studies will thus be adopted subsequently, in relation to the discussion and comparison of the results presented in the tables, and in the qualitative analysis of particular word-combinations in chapter 5.

4.3.1 Full Clauses and Multiple Clause Constituents

Tables 4.9 and 4.10 show the full clauses and multiple clause constituents in the LOCNEC and LINDSEI-SW material which consist of a pronoun in subject position, including all uses of *it* and the existential *there*, plus one or more additional items. Combinations including the same pronoun are listed together and chronologically according to frequency, and separated by a space. The combinations which consist of a pronoun and other items, but no verb, e.g. *and I*, are listed separately. All the combinations that can be embedded into other longer combinations, e.g. *I mean*, are underlined, and the longer combinations are italicized and ordered according to frequency. This form-based classification makes the combinations and their frequencies easier to access and compare, and highlights the fact that many of the smaller combinations are embedded into longer combinations which are also recurrent enough to make it past the set frequency thresholds.

LOCNEC:
FULL CLAUSES AND MULTIPLE CLAUSE CONSTITUENTS

Pronoun + verb phrase	
444[15]	**I mean**
84	*I mean I*
7	*I mean I was*
7	*but I mean I*
34	*but I mean*
34	*yeah I mean*
25	*I mean it's*
11	*know what I mean*
437	**I was**
56	*and I was*
7	*and I was just*
47	*when I was*
29	*I was there*
11	I was going to
433	**I think**
53	*I think I*
52	*I think it's*
7	*I think it's a*
38	*I think it*
19	*I think it was*
33	*but I think*
28	*and I think*
25	*yeah I think*
423	**I don't**
209	*I don't know*
53	*I don't know I*
17	*I don't know if*
10	*but I don't know*
9	*I don't know it*
9	*I don't know what*
7	*well I don't know*
46	*I don't think*
7	*I don't think I*
7	*I don't think so*
25	*but I don't*
9	*I don't really know*
9	*I don't want to*
138	**I went**
56	*I went to*
7	*when I went to*
11	*I went to see*
9	*and then I went*
127	**I had**
45	*I had to*
8	*I had to go*
7	*so I had to*
120	**I thought**
13	*I thought it was*
9	*I thought that was*
98	**I'm not**
94	**I did**
90	**I've got**
86	**I know**
84	**I didn't**
79	**I can't**
62	**I'd like to**
19	*I'd like to go*
8	*I'd like to teach*
38	**I wanted to**
9	*I wanted to do*
7	*what I wanted to*
31	**I want to**
12	*I want to do*
8	*what I want to*
25	**I used to**
7	**I didn't want to**
757	**it was**
90	*and it was*
79	*it was a*
11	*it was a bit*
8	*it was a lot*
7	*so it was a*
66	*it was just*
11	*and it was just*
11	*it was just like*
52	*it was really*
17	*it was really good*
13	*and it was really*
47	*it was it*
45	*yeah it was*
9	*yeah yeah it was*
40	*it was like*
12	*and it was like*
7	*it was like a*
39	*it was quite*
7	*it was quite a*
7	*it was quite good*

[15] All numbers refer to raw frequencies.

→		Pronoun + conjunction/response item	
38	*so it was*	**208**	**yeah I**
34	*it was very*	**367**	**and I**
27	*but it was*	**239**	**but I**
147	**it's not**	**224**	**so I**
9	*it's not too bad*	**137**	**when I**
142	**it's a**	**125**	**well I**
23	*it's a bit*	**102**	**I just**
126	**and it's**	**101**	**that I**
122	**it is**	**100**	**no I**
111	**it's just**	**87**	**cos I**
90	**yeah it's**	**80**	**if I**
9	**it would have been**	**49**	**and then I**
632	**you know**	**174**	**and it**
43	*you know you*	**101**	**of it**
41	*you know I*	**90**	**yeah it**
38	*and you know*	**82**	**but it**
33	*you know it's*	**78**	**so it**
23	*you know the*		
12	*you know what I*	**166**	**and you**
10	*you know it was*	**104**	**if you**
150	**you can**	**137**	**and they**
86	**you have**	**101**	**and he**
56	*you have to*	**92**	**and we**
9	*and you have to*	**86**	**and she**
7	*you have to go*		
201	**that was**		
28	*so that was*		
26	*and that was*		
23	*that was a*		
83	**that's right**		
35	*that's right yeah*		
26	*yeah that's right*		
7	*yeah that's right yeah*		
102	**we were**		
86	**we went**		
30	*we went to*		
7	*we went to the*		
101	**there was**		
26	*there was a*		
94	**they were**		

Table 4.9: *LOCNEC: Full clauses and multiple clause constituents, 2-4-word combinations (freq. >77/23/7)*

Table 4.9 presents the data from LOCNEC, and emphasizes how certain pronouns, verbs and conjunctions frequently re-occur in the spoken language of native speakers. The table reaffirms the impression from the top 20-lists above that the *first person singular pronoun + verb*-combinations are very common in the corpus, which is largely to be expected partly as a cause of the text- and task types, where the speaker is urged to talk about him/herself. This tendency is also reflected in the many *pronoun + conjunction/response item* combinations with *I* occurring in the corpus, e.g. *and I*, *well I*, as seen in the bottom right column. Table 4.9 thus corresponds with Biber et al.'s (1999) observation from their conversational data, that "most of the sequences made up of the following elements occur as recurrent lexical bundles in conversation: *I/you + don't/didn't + know/think/want + complement-clause*" (Biber et al. 1999: 1000). In addition, table 4.9 shows how most of the 3- and 4-word combinations in the data include a highly frequent 2-word combination. Some of these 2-word combinations, like *I don't* (n=423), seem to be high up on the frequency list mainly due to the high frequencies of longer combinations, like *I don't know* (n=209) and *I don't think* (n=46), which together make up more than half of the instances of *I don't*. Other combinations with *I don't* include *I don't want* (n=11), representing the last of the three verbs in Biber et al.'s listing of common recurrent consecutive elements. In another example of embedding, the recurrence of the 3-word combination *you have to* accounts for more than half of the occurrences of the frequent *you have* in LOCNEC. This combination seems to occur mainly as a result of the interview situation, where the subject has to explain concepts and procedures (7):

(7) *<A> oh so you already have to teach <\A>*

<B> oh you take over the class from the teacher and you have to plan everything <\B> (LOCNEC)

4.3.1.1 Versatile word-combinations in LOCNEC: I mean, I think, you know

Although many 2-word combinations can thus be embedded into longer combinations, the frequency figures in table 4.9 also suggest that combinations like *I mean*, *I was* and *you know*, which were found to be significantly underused by the Swedish and Norwegian learners in section 4.2.1, are not highly frequent in LOCNEC mainly as a result of embedding, but from being individually frequent in a range of linguistic contexts, exemplified in 8-10:

(8) *yeah .. I mean there'd there'd been a drive by shooting .. erm where I was staying the week before* (LOCNEC)

(9) *I think it's a nice campus and it's not I mean it could be a lot worse you could be surrounded by: lots of . at least you've got the trees the duck pond* (LOCNEC)

(10) *em well no er I mean I'm I'm doing biology because that's . it's the one subject I've I've always found easy and I enjoy it* (LOCNEC)

In (10), *I mean* is followed by the contracted form *I'm*, which is treated as one word in the corpus software used to retrieve frequency data in this study. The frequency numbers for combinations ending with a pronoun, such as *I mean I*, *I think I, I don't know I* and *you know you/I*, are thus likely to have been even higher if contracted forms were split up[16]. The frequency of these 3-word combinations with a final pronoun also suggests that the 2-word combinations that are embedded in them (*I mean*, *I think, I don't know* and *you know*), although they are contextually and functionally versatile, often appear as stems in front of clauses of personal reference, or at the beginning of clauses as frames consisting of "'thematic elements' in pre-subject position" (Altenberg 1998: 112). The frame position is seen in (10), where *I mean* occurs with two filled pauses and the initiator *well*. In (11), however, it seems unclear whether *I think* functions as an epistemic stem modifying the following content, or as a frame, similar to *I mean*, and in combination with *so er* preceding it:

(11) *yeah rather than specialise so er I think I'm quite happy with what I've chosen* (LOCNEC)

I mean and *you know* are classified by Biber et al. (1999) as discourse markers, "inserts which tend to occur at the beginning of a turn or utterance" and which "combine two roles: (a) to signal a transition in the evolving progress of the conversation, and (b) to signal an interactive relationship between speaker, hearer, and message" (Biber et al. 1999: 1086). This interactive function seems to be prevalent also in the LOCNEC data, although there are examples in the material where you know functions declaratively as a main clause followed by a *that*-clause (with or without the subordinating conjunction), rather than as an interactive

[16] A search for *I mean I'* in LOCNEC, using a simple text editor, retrieved 28 instances, with *I've* (n=12), *I'm* (n=11), *I'd* (n=3) and *I'll* (n=2) as the contracted forms.

discourse marker (12-13), thus illustrating that even well known word-combinations with discourse functions may display meanings which are closer to the semantic meaning of their component parts.

(12) *yeah when you go on it's like when go on holiday abroad you when you know that pubs are open are open until five in the morning or whatever you . you tend to sort of grab for a meal maybe about nine o'clock* (LOCNEC)

(13) *I mean .. many people have said this that you know you wanna become an actor why don't you do the theatre studies* (LOCNEC)

I think, on the other hand, is classified as an 'epistemic stance adverbial' by Biber et al., "used to present speaker comments on the status of information in a proposition" (Biber et al. 1999: 972) rather than a discourse marker, in agreement with Altenberg's (1998) definition of epistemic stems. This classification may also apply to frequent combinations like *I don't know*, although both *I think* and *I don't know* seem to perform functions beyond the marking of epistemic stance, as seen in (14) and (15), where these combinations appear in combination with other discourse markers like *I mean* and *you know*, and seem to function as planners in a frame position, in addition to the epistemic revision of the following content:

(14) *and er it was I think you know the way that they sort of the drummers have<'> used all these really strange percussion items as well* (LOCNEC)

(15) *I don't know I might I might have to .. yeah I think I mean sometimes you have to do like a an accelerated year of just teaching .. erm . like just like a teaching course really* (LOCNEC)

Discourse functions of word-combinations with *I think* and *I don't know* are found in other studies of spoken English (Aijmer 1997, 2004; Baumgarten and House 2010; Kärkkainen 2003), and this will be discussed further in relation to the comparison with the learner material.

The high frequencies of the word-combinations discussed above is in itself an indication of versatile usage patterns, and the frequency numbers retrieved for their most common collocations confirm this assumption. The retrieval of recurrent word-combinations and the recurrent sequences in which they are embedded may thus provide indications of how frequent word-combinations typically function, and in what linguistic contexts they typically appear, although further qualitative analysis is needed in order to confirm any assumptions made. This quantitative information also

provides a basis for a comparison between the material from LINDSEI-SW. The corresponding data on full clauses and multiple clause constituents in the Swedish LINDSEI is presented in table 4.10:

Pronoun + verb phrase

533 I think
60 I think it's
7 I think it's a
5 and I think it's
5 so I think it's
5 but I think it's
5 I think it's more
51 I think it
9 I think it is
23 I think it was
23 think it was
46 I think I
6 I think I would
5 no I think I
38 I think so
19 yeah/yes I think so
30 but I think
28 I think that's
6 I think that's the
6 and I think that's
27 and I think
25 yeah I think
21 I think that
7 I think that was
19 I think the
18 so I think
17 I think they
16 I think and
15 no I think

300 I don't
139 I don't know
(81 I dunno)
20 I don't know if
10 I don't know what
9 so I don't know
8 I don't know really
8 I don't know how
7 I don't know I
7 I don't know why
5 and I don't know
57 I don't think
16 I don't think so
10 no I don't think
6 I don't think I
5 I don't think they
24 no I don't
17 but I don't
16 so I don't

204 I was
31 when I was
6 when I was there
5 when I was a
5 when I was younger
25 I was there
15 and I was

167 I mean
29 I mean I
16 I mean it's
15 but I mean
5 I mean they have

104 I like
18 I like the

81 I have
76 I didn't

76 I'm not
22 I'm not sure

72 I would
20 I would say
17 I would like
15 I would like to

71 I had
22 I had to
7 so I had to
62 I haven't
59 I guess
51 I do

50 I went
5 before I went there
21 I want to
20 I went to
6 I thought it was

419 it was
51 and it was
37 it was very
6 so it was very
6 and it was very
5 but it was very
6 it was very nice
34 it was a
7 and it was a
27 but it was
23 yeah it was
22 it was just
21 so it was
20 it was like
5 it was like a

→		Pronoun + conjunction/response item	
122	**it's a**		
95	**it's not**	**236**	**and I**
80	**it is**	**170**	**but I**
59	**but it's**	**168**	**so I**
57	**so it's**	**126**	**yeah I**
53	**and it's**	**85**	**when I**
19	**it would be**	**81**	**no I**
6	**it's hard to say**	**79**	**that I**
		68	**well I**
204	**you know**	**63**	**I just**
20	*you know the*	**58**	**if I**
17	*and you know*	**57**	**because I**
72	**you can**	**50**	**then I**
70	**you have**		
40	*you have to*	**79**	**and it**
	5 you have to go	**51**	**it and**
17	**you want to**	**51**	**but it**
8	**what do you say**	**51**	**but it**
8	**if you want to**		
		89	**if you**
109	**that was**	*5*	*if you look at*
17	*that was a*	**77**	**and you**
	6 that was a good	**51**	**when you**
18	*so that was*		
18	*and that was*	**92**	**and they**
5	**that would be interesting**	**65**	**and she**
		60	**and he**
51	**we went**	**54**	**and we**
21	*we went to*	**50**	**so we**
	5 we went to the		
5	*and then we went*		
93	**they were**		
77	**they have**		

Table 4.10: *LINDSEI: Full clauses and multiple clause constituents, 2-4-word combinations (freq. >50/15/5)*

The LINDSEI table, like the LOCNEC-table, shows a dominating presence of the first person pronoun, a limited number of highly frequent verbs (conjugations of *think, do, be, mean, like, will, can, have, guess, go, want* and *know*), and several co-ordinating and subordinating conjunctions. The extended material also strengthens the impression that the spoken recurrent language of Swedish learners of English is very similar to the native speaker norm, as it is difficult to spot differences in combination types in the two tables. However, some differences do occur, and these differences might serve as potential markers of non-nativeness in the data.

4.3.1.2 I think

Unsurprisingly, table 4.10 comprises many different combinations in which the overused combination *I think* is embedded. One of the most notable combinations is perhaps the 3-word combination *I think so* (n=38), which typically occurs as a response to a direct question (16) or a declarative clause functioning as a question (17):

(16) *<A> and can you get that combination English and history </A>*

<B> yeah I think so </B>

<A> (mhm) </A>

<B> it's very .. they need (eh) teachers now <overlap /> so it </B>

<A> <overlap /> (mhm) </A>

<B> so I don't think it will be any problem <overlap /> to get </B>

<A> <overlap /> no </A>

<B> a job so </B> (LINDSEI-SW)

(17) *<A> (uhu) .. are you a little bit more (em) <tuts>. sceptical now about a relationship or do you think that </A>*

<B> yeah I'm ver= very much so I haven't been in a <breathes> long relationship . since </B>

<A> (uhu) . so you're more demanding now </A>

<B> I think so </B> (LINDSEI-SW)

This combination has not made it onto the LOCNEC list (with its >23 threshold), and the high frequency in the learner data thus contributes considerably to the overall pattern of overuse of *I think*. Similar findings are reported by Baumgarten and House (2010) in their study of conversational data from German speakers of English as a lingua franca, where *I think so* is recurrent in the non-native speaker data, while at the same time being absent in the data from their control group of native English speakers (Baumgarten and House 2010: 1190). *I think* thus dominates table 4.10 because of its sheer frequency, but also because it is embedded into a number of recurring word-combinations. To further illustrate how dominant *I think* is in both corpora, table 4.11

shows the words most commonly collocating with *think*, in both left and right position[17]:

	LOCNEC			LINDSEI-SW		
n	Fr.	Left	Right	Fr.	Left	Right
think	**574**			**666**		
I		**433**	60		**533**	48
don't		49	0		64	0
to		14	1		2	1
it's		0	58		0	76
it		0	42		0	56
I'll/I'm/I'd/I've		3	39		0	14
they/you		27	28		19	49
that		0	23		0	37
the/this		0	22		0	22
so		0	19		0	50
that's		0	13		0	30
they're/they've/they'd		0	13		0	3
we		1	12		0	9
if		0	11		0	2
about		0	11		0	10
oh		0	10		0	2
of		0	9		2	6
is		0	8		0	4
and/but/well		6	8		0	12
er/erm/em/eh		0	8		1	11
she		0	7		0	15
yeah		0	5		0	4
	Σ	533	407	Σ	621	461

Table 4.11: *<u>think</u> and some of its most frequent collocational patterns in LOCNEC and LINDSEI-SW, with raw frequencies*

I think proves to be not only a very common combination as compared to other word-combinations in both corpora, the two words also show an overwhelming tendency to co-occur. In LOCNEC, *think* collocates with *I* in initial position 75.4 % of the times it occurs, while the corresponding percentage for LINDSEI-SW is 80 %. In addition, the second most frequent collocation in this position is *don't,* which almost exclusively takes part in the significantly overused combination *I don't think* in both corpora (see table 4.12).

[17] Some related collocates, like *they* and *you*, are manually conflated to make the table more readable.

LOCNEC		LINDSEI-SW		
n	***n* per 10,000**	***n***	***n* per 10,000**	**X^2**
46	3.9	57	7.9	**13.24**

Table 4.12: *LOCNEC&LINDSEI-SW: I don't think, absolute frequencies, n per 10,000 and chi-square result (d.f. = 1)*

Most of the collocations in the right hand columns are thus collocating with *I think* or *I don't think*, such as the above mentioned *so. That* is more commonly collocated with *think* and *I think* in the learner corpus compared with the native speaker corpus (*I think that* occurs 1.4 times per 10,000 words in LOCNEC, 2.9 times in LINDSEI-SW), and seems to function most often as a subordinating conjunction in LINDSEI-SW (18), but also as a referential pronoun (19):

(18) *and (eh) I think that I would like to go to Athens and see .. you know Acropolis and <sighs>* (LINDSEI-SW)

(19) *<starts laughing> and I think that <stops laughing> was really interesting* (LINDSEI-SW)

That also functions as a referential pronoun in the frequent 3-word combination *I think that's* (20), although this combination also appears in contexts where the antecedent of *that* is not clearly visible, and the clause is interrupted and restarted (21):

(20) *so (erm) . it was not really too far away just two days and that's . yeah I think that's okay <laughs>* (LINDSEI-SW)

(21) *<B> I think I think that's (eh) there isn't as much tourism in Cuba <overlap /> as it </B>*

<A> <overlap /> (mm) </A>

<B> is in the Dominican Republic <breathes> and therefore <breathes> maybe people do <breathes> they have a bit m= more money because they make money off th= the tourists maybe </B> (LINDSEI-SW)

Similar behaviour can be seen with *I think it's*, which is also overused in LINDSEI-SW (occurring 8.4 times per 10,000 words, as opposed to 4.4 times in LOCNEC), where *it* can be anticipatory and express predicative relations (22), or emptier of reference and in a context of hesitation (23):

(22) *yeah but I mean I think it's hard to be as good in two languages* (LINDSEI-SW)

(23) *but I think it's more (eh) it's cheaper there bec= than (em) . on the big islands because it's n= not man= not as many tourists there* (LINDSEI-SW)

The 4-word combination *I think it's more* in (23) is also overused by the Swedish learners, with five occurrences altogether, as opposed to two in LOCNEC. The high frequency of *I think it's* and *I think that's,* and the hesitation and interruption surrounding them in the beginning of utterances could signify holistic storage of these 3-word combinations, leading to easy retrieval in situations where the learners experience difficulties in expressing lexical or propositional content. This usage pattern is also found in the native speaker corpus (example 24), which indicates that the possible holistic treatment of this 3-word combination could be based on input from native language:

(24) *no they wouldn't I think it's just . you have nerv= you're nervous and* (LOCNEC)

The suggestion that *I think it's* and *I think that's* functions formulaically is thus supported by their high frequency of occurrence (both combinations occur on the LINDSEI-SW top 20-list of 3-word combinations in section 4.1), combined with evidence of use in discourse frames followed by markers of hesitation.

Aijmer (1997) describes the different meanings of *think* in terms of a polysemic structure, with cogitation as the focus of the structure from which the meanings of belief, opinion and intention derive (Aijmer 1997: 12). The strictly prototypical meanings of *I think* as referring to cognition seem to be rare, but some instances can be found with the 3-word combination *I think of* in both the native speaker and the learner corpus:

(25) *and then I get very angry because I think of . (er) how they will feel themselves when (eh) if they hit somebody if they kill somebody* (LINDSEI-SW)

(26) *yeah so <X> possibly I think of going out there for a few years of teaching English* (LOCNEC)

The expression of opinion and belief can sometimes be hard to distinguish in actual language use, but some instances of *I think* can be separated as being tied closer to belief rather than opinion:

(27) *and (eh) that's what I think I'll be doing (eh) teaching French English at high school* (LINDSEI-SW)

(28) *but I think I'll mainly be doing exams anyway* (LOCNEC)

This distinction can be made clearer by means of translating into Swedish or Norwegian where the belief-sense of *I think* corresponds to *jag/jeg tror*, and the opinion-sense corresponds to *jag/jeg tycker/synes* (Aijmer 1997: 15).

In her study of epistemic stance in American English conversation, Kärkkainen (2003) argues for the description of *I think* as mainly a discourse marker, rather than a compositional phrase referring to cognition, or an adverbial expressing epistemic meaning, as in Biber et al.'s (1999) classification. Kärkkainen finds that many of the occurrences of *I think* in her native speaker data lack referential meaning, and thus confirm a "strengthening of conversational implicatures and (...) development of new pragmatic meanings" (Kärkkainen 2003: 178). Through this development, *I think* "has not completely lost all semantic meaning but this meaning has become latent, waiting to be put to use when needed" (ibid.). One of the functions of *I think* as a discourse marker is that of "a starting-point function" (Kärkkainnen 2003: 179), which thus positions the word-combination in Altenberg's (1998) discourse-organizational frame. As seen above, this usage is also found with longer combinations in the learner data, such as *I think that's* and *I think it's*, suggesting that these word-combinations might be subject to some of the same processes.

Aijmer explains the extended meanings of *I think* in native speaker English according to a "cline of pragmaticalization" (Aijmer 1997: 6), which is separated from *grammaticalization* in that pragmaticalized elements are typically optional in the clause, whereas constructions derived from grammaticalization are obligatory as markers of mood or tense (e.g. *going to* expressing future tense). The first stage of this process is the move from the transparent meaning of 'cogitation' towards those of opinion and belief, through a metaphoric strategy: "speakers view the formation of an opinion or belief in terms of thinking and borrow the verb *think* to express the new meanings" (ibid.: 12). Moreover, Aijmer supposes that pragmaticalization "is a dynamic and fluctuating phenomenon" (ibid.: 40), and the fact that *I think* is still compositional and used as a main clause strengthens this assumption. However, *I think* seems to have "developed a number of new functions as a response to the

demands of planning and interaction with the hearer" (ibid.), such as the use of *I think* to soften speech and "avoid bluntness" (ibid.: 20).

Since many of the same usage patterns are found in both LOCNEC and LINDSEI-SW, no single explanation of the highly significant overuse of *I think* has thus far been provided. It is possible that the native speaker use of *I think* is, in accordance with Kärkkainen's findings, mainly discourse oriented, whereas the Swedish learners make use of the combination in a larger variety of context, including those of a more proposition-related kind, and those expressing epistemic stance. This is supported e.g. by the larger frequency of *I think so* and *I think that* in the learner corpus. Considering the use of *I think* as a discourse marker typically functioning in the frame of an utterance, it is likely that learners would also have a greater need for this sort of item, since their processing difficulties are more prominent. Another explanation for the overuse of this combination in all contexts might be related to a general tendency among learners to express doubt about propositions made, which will be further discussed in the following sections.

4.3.1.3 I don't know/I dunno

An interesting find in table 4.10 is the combination *I dunno*, which, when added to the frequency count for *I don't know*, makes this figure considerably higher, as shown in table 4.13. This combined frequency count thus changes the statistics presented in table 4.8, where the difference in frequency of *I don't know* between the NS and the Swedish NNS corpora was not found to be statistically significant.

LOCNEC		LINDSEI-SW		
n	***n* per 10,000**	***n***	***n* per 10,000**	**X^2**
209	17.8	139+81=220	60.9	**32.47**

Table 4.13: *LOCNEC&LINDSEI-SW: I don't know + I dunno, absolute frequencies,* n *per 10,000 and chi-square result (d.f. = 1)*

The non-standard form *dunno* cannot be found in the LOCNEC corpus, which suggests that it was not part of the transcription conventions for this corpus. The reduced form of *don't* in recurrent word-combinations and its relation to differences in discourse functions is analysed by e.g. Scheibmann (2000), and the phonetic

reduction of high-frequency words and word-combinations, with particular reference to *I don't know*, is investigated by Bybee (2010), who claim that "in general the bias towards reduction is a result of chunking: as sequences of units are repeated the articulatory gestures used tend to reduce and overlap" (Bybee 2010: 37). The appearance of *I dunno* in LINDSEI-SW shows a presence of this reduced form and indicates a process of 'chunking' of this combination in the learner speech production, but since there is no basis for comparison with LOCNEC, and since phonological information is not considered for any of the other combinations in the data, *I don't know* and *I dunno* are considered the same word-combination for the purpose of this study.

The learner data, like the native speaker data, shows that the combination *I don't* (n=300) is mainly embedded into *I don't know* (n=139) and *I don't think* (n=57). One of the 4-word combinations which further contributes to the frequency of *I don't know* in the learner corpus*, I don't know what* (n=10), may represent a particular use of *I don't know* mainly attributed to learner language, where the combination "takes its literal meaning" (Aijmer 2009: 154), and where learners experience that "the difficulty of expressing themselves in a foreign language interferes with the encoding process as a whole" (De Cock 2004: 234) (29-30). Retrieval difficulties are also found to be signalled through *I don't know* only (31):

(29) *<overlap /> (eh) jo= <foreign> johanniterordern </foreign> it's in Swedish <overlap /> I don't know what it is in English* (LINDSEI-SW)

(30) *because he's pumping money into the business <breathes> so: and she's she's ill as well she has a: I don't know what you call it (eh) . in the lungs* (LINDSEI-SW)

(31) *<B> and (er) I don't know . <foreign> konjunktiv </foreign> </B>*

<A> okay . yeah </A

<B> I really don't know what it's called in S= English </B> (LINDSEI-SW)

This usage is reported by Götz and Schilk (2011) to be the most common function of *I don't know* in the German component of LINDSEI, signalling "uncertainty about the linguistic features of a speakers' utterances" (Götz and Schilk 2011: 93). However, the lexical retrieval problems seen in examples (29)-(31) only occurs with four

instances of *I don't know what* in LINDSEI-SW, and, interestingly, there are no instances of *I don't know how to say/I don't know how you say* as an overt lexical retrieval strategy in the corpus, unlike in the French LINDSEI, where these are common word-combinations used for this purpose (De Cock 2004: 234). The 4-word combination *what do you say* (n=8) in table 4.10 does however take on this function (example 32), and this combination occurs once for the same purpose in LOCNEC (example 33). The higher frequency in LINDSEI-SW is thus a further indication of greater lexical retrieval difficulties:

(32) *<B> <XX> yeah I think so . people are more (mm) suspicious and you know . closed or <u>what do you say</u> <laughs> more </B>*

<A> right </A>

<B> snobbish maybe I don't know </B> (LINDSEI-SW)

(33) *it is .. well there are people who go and like I know someone who goes to Caton every Friday to help with the scout . erm pack or <u>what do you say</u> troop scout troop and there's someone else who goes to Galgate .. and helps with one we've had people who ran guide <X> erm .. guide packs* (LOCNEC)

Returning to the overused *I don't know/I dunno*, table 4.10 indicates that both of these combinations appear in a variety of contexts. The combination *I don't know I* is not particularly frequent in the non-native speaker corpus, with 7 occurrences as opposed to 53 in the LOCNEC corpus. If contracted forms such as *I don't know I'm* are included, the frequency increases to 12 in LINDSEI-SW and 60 in LOCNEC. It was suggested above that these personal pronoun-combinations might indicate that the embedded combination often occurs at the beginning of an utterance, as either a 'discourse marking' frame or as an epistemic stem evaluating clauses with propositional content which is subject to some sort of modification by the speaker. The comparatively low frequency of *I don't know I* in LINDSEI-SW thus suggests that this is not a common position for *I don't know* in the discourse in the learner corpus, although it is not possible to make firm claims on the basis of data of such limited context. Examples 34 and 35 show the frame position of *I don't know*, in combination with response items and filled and unfilled pauses:

(34) *now but no . <u>I don't know</u> you just feels so spoilt here in Sweden and I like that* (LINDSEI-SW)

(35) *<A> would you like to paint like that yourself . in that way </A>*

<B> (er) .. yeah pff I dunno I don't paint <laughs> </B> (LINDSEI-SW)

In the recurrent word-combinations where *I don't know* function as a main clause, the more transparent meaning of cognition, 'not knowing', is more obviously linked to the subsequent complement clause, and the speaker is showing his or her insufficient knowledge:

(36) *okay . (mm) <sniffs> well there's a man there a painter or I don't know if he's a <starts laughing> professional but <stops laughing>* (LINDSEI-SW)

(37) *yeah they do . I don't know why they why they can't have real people like the way they <starts laughing> really look <X> I mean <laughs>* (LINDSEI-SW)

These combinations are more common in the NNS corpus, with 6.3 occurrences per 10,000 of *I don't know if/what/how/why* as compared to 3.1 per 10,000 in LOCNEC. It is possible that this use of *I don't know* is more common in the learner corpus, and that this is linked to an extended fear of being assertive about propositions made. Baumgarten and House (2010) find that their German ELF speakers are more oriented towards themselves in speech, and that they more often than native speakers make use of the 'prototypical' *I don't know,* which indicates lack of knowledge (Baumgarten and House 2010: 1198). This assumption is compatible with the finding that *I don't know* functioning as a main clause with a complement seems to be more common in LINDSEI-SW, based on its most common collocations. In addition, this contrast between a more literal, proposition based use of *I don't know* in learner speech, and a more interactive, discourse marking function of the native speaker *I don't know*, agrees with the assumptions made in relation to the usage patterns of *I think* in section 4.3.1.2 above. Nevertheless, there seems to be a complex and varied use of both *I think* and *I don't know* in both learner and native speech, which is reflected in their high frequencies and the linguistic context they appear in. In addition, the significantly higher frequency of both of these combinations in the learner corpus may indicate an overuse of these combinations functioning in both interactive and discourse related contexts *as well as* those modifying propositional content. Possible reasons for the overuse of these word-combinations will be discussed further in chapter 5.

4.3.1.4 Certainty and doubt: that's right, I guess

While there are many similarities between the recurrent word-combination types including pronouns and verbs in the two corpora, such as *I think* and *I don't know*, the 2-word combination *that's right*, which occurs 83 times in the LOCNEC corpus, is notably absent from the LINDSEI table. A search in the LINDSEI-SW corpus reveals that this combination only occurs once in the data as a whole, indicating a significant underuse, which is also reported by De Cock (2004) in relation to the French component of LINDSEI (De Cock 2004: 242). It is possible that this underuse is also related to a greater learner uncertainty or fear of being too assertive in learner speech, and that other overused combinations like *(yeah) I think (so)* or even *I don't know*, is used in its place, as "an uncertainty device" (Aijmer 2004:188). Contrary to this assumption, the Swedish learners' underuse of *that's right* might also be related to the possible overuse and/or misuse of the 2-word combination *of course*, as reported in De Cock (2004), which will be further looked into in relation to the single clause constituents in section 4.3.2 below.

In LOCNEC *that's right* seems to be used mainly to confirm the previous statement from the interviewer (38), while *I think so* is found to be used in a similar fashion in LINDSEI-SW (39):

(38) *<A> mhm and the carnival with it's in Venice isn't it <\A>*

<B> yeah <\B>

<A> with all the masks <\A>

<B> <u>that's right</u> yeah yeah we went there when that was on but er we just went round the museums <\B> (LOCNEC)

(39) *<A> (mm) ... it wasn't about being gay and the problems <overlap /> of living <?> </A>*

<B> <overlap /> no </B>

<A> no it wasn't </A>

<B> more like being different or .. </B>

<A> just generally different </A>

<B> <u>I think so</u> </B> (LINDSEI-SW)

Baumgarten and House (2010) explain many of the occurrences of *I think* produced by the non-native speakers in their study on the basis of this awareness of stance-taking in relation to propositions made, and that "the expression of subjectivity is seen as potential trouble spot" (Baumgarten and House 2010: 1197) for the non-native speakers, in terms of being too assertive about their opinions. Related to this is the occurrence of *I guess* (n=59) in LINDSEI-SW, which does not occur in the LOCNEC table, and can only be found 11 times in total in the NS corpus. This discrepancy indicates that *I guess* is overused by the Swedish learners, which adds to the general impression of tentativeness in the learner data which the patterns of overuse of e.g. *I think* and *I don't know* generate. *I guess* occurs 21 times in the Norwegian sample, which makes for a highly significant overuse of this combination in both learner populations, as seen in table 4.14.

LOCNEC		LINDSEI-SW		
n	***n* per 10,000**	***n***	***n* per 10,000**	**X^2**
11	0.9	59	8.2	63.85
		LINDSEI-NO		
		n	***n* per 10,000**	**X^2**
		21	5.8	31.34

Table 4.14: *LOCNEC, LINDSEI-SW and LINDSEI-NO: I guess, absolute frequencies, n per 10,000 and chi-square result (d.f. = 1)*

Table 4.10 does not reveal whether *I guess* occurs as part of longer recurrent word-combinations, but a search for the response *I guess so* reveals one occurrence in LINDSEI-NO and three in LINDSEI-SW. The general overuse of *I guess* might be predominantly due to the considerable input from American English Swedish and Norwegian students normally receive through media, as noted by e.g. Aijmer (2004:185) (see also table 3.3 on the impact influence from the media has on the Swedish learners' proficiency). A search in the spoken component of the Corpus of contemporary American English (COCA), which consists of transcripts of unscripted conversation from TV and radio programs, retrieves 236.92 occurrences of *I guess* per one million words, whereas the spoken part of the British National Corpus (BNC) confirms the findings from LOCNEC, with only 16.16 occurrences of *I guess* per one million words. This influence from American English is also thought to be relevant for the overuse of *kind of* (and the corresponding underuse of similar combinations

like *sort of*) among learners (Aijmer 2004; De Cock 2004) (see section 4.3.2. below), and might also be relevant for the analysis of other deviating patterns in the LOCNEC/LINDSEI comparison which are more difficult to detect. However, as mentioned, the extended use of *I guess* in spoken learner language seems to be falling into a pattern of an extended use of markers of hesitation in general, which is perhaps part of the reason why this particular word-combination is favoured by the Swedish and Norwegian learners. A more in-depth comparison with COCA, which seems to be a relevant representation of the American English input Scandinavian learners typically receive, would reveal whether this usage pattern also deviates from the American native speaker norm and might thus confirm this impression, an undertaking which is beyond the scope of this study.

4.3.1.5 Highly significant patterns of underuse: you know *and* I mean

The underuse of *you know*, which was found to be highly significant for both LINDSEI-SW and LINDSEI-NO in table 4.8/section 4.2.1 above, might be another trait related to a wish of being less assertive in learner language. *You know* in native speech is, as mentioned above, often used as a discourse marker which signals "an interactive relationship between speaker, hearer, and message" (Biber et al. 1999: 1086) and, more specifically, acts "as a 'shared knowledge indicator' signaling the speaker's confidence in the existence of common information" (House 2009: 172). House (2009) found in her study of *you know* in ELF (English as a Lingua Franca) that the non-native speakers rather used the word-combination as a "self-serving strategy" (ibid.: 178), and as a pragmatic device mainly "to monitor their own progression in discourse" (ibid.: 189), through planning utterances and connecting propositions. An example of this can be seen in (40), where the Swedish learner "cannot find the right words, fumbles for the appropriate word or formulation" (House 2009: 186), and uses *you know*, perhaps to reveal this difficulty to the hearer:

(40) *yeah it is . but I haven't been to (mm:) you know the <foreign> ja gymnasiet </foreign>* (LINDSEI-SW)

In LOCNEC, an example of speaker-hearer involvement expressed by *you know* becomes apparent when *you know* occurs in the 4-word combination *you know what I*, which in turn is embedded in the 5-word combination *you know what I mean* in all its 12 occurrences in the native speaker corpus:

(41) *because like .. I seem to be a bit sort of .. I wanted to sort of get away from living at home if you know what I mean* (LOCNEC)

Judging from table 4.10 this combination is not as frequent in LINDSEI-SW, and searching the corpus only retrieves two occurrences. House (2009) suggests that the non-interactive functions of the learners' use of *you know* is due to its formulaic nature, where its meaning has become pragmaticalized and detached of literal meaning:

> *"You know* is used in a highly conventionalized way [in non-native speech], which means that the original meaning of *you know* is no longer virulent, it is but a stock phrase mainly used to help speakers process and plan their output, and link spans of discourse. It has little to do with hearer deixis or a second person perspective: as a formula it does not enter into the consciousness of interactants implying mutual engagement and participation" (House 2009: 189).

In this perspective, *you know* does not fully function as a discourse marker in the sense defined by Biber et al. "to signal an interactive relationship between speaker, hearer, and message" (Biber et al. 1999: 1086), since the hearer is somewhat left out of the equation. Instead, the combination is used to promote the needs of the non-native speaker in his or her production: "As an instance of formulaic language *you know* is fully functional— primarily (...) for the benefit of the speaker him or herself" (House 2009: 190). This is also a possible interpretation of the use of *you know* in (42):

(42) *we didn't see many: you know historical sights I don't know if there were any <laughs>* (LINDSEI-SW)

If this functional pattern is valid also for learner English (and not only peculiar to ELF), it does not fully explain the highly significant pattern of *underuse* revealed in LINDSEI-SW, since such 'highly conventionalized' combinations are likely to be very frequent. However, if *you know* is predominantly used for 'self-serving' purposes in the learner speech, and the other more literal uses are disregarded either consciously, to avoid making assumptions about knowledge on behalf of the interviewer, or unconsciously, as a result of the pragmaticalization of the word-combination, this might serve to partly explain why the combination is used more often by the native speakers, who might thus use it for a wider range of purposes. However, it is also possible that both the propositional and interactive meanings of *you know* are indeed present in the learner minds, which in turn leads to the rejection

of the combination because of its assumed function as a 'shared knowledge indicator'. Furthermore, the underuse of *you know* might be related to the limited reference to the second person learners seem to make in general, considering the fact that *you* occurs 154 times per 10,000 words in LINDSEI-SW as opposed to 196 times in LOCNEC. Since *you know* is classified as a discourse marker it should perhaps not be associated too closely with the meaning of its component parts, but the high frequency of the combination *you know you* in LOCNEC (and a correspondingly low frequency of n=3 in LINDSEI-SW) indicates that there might be a connection here.

I mean, also a frequent discourse marker in native English speech, has been found to be significantly underused by the Swedish and Norwegian learners, but the combination is still one of the most frequent combinations in LINDSEI-SW. Some of the longer combinations in which *I mean* is embedded, such as *I mean I, I mean it's* and *but I mean*, are frequent in both the native and non-native speaker corpora. It was suggested above that, judging from its collocational patterns, *I mean* is a very versatile combination, which seems to be the case also of *I mean* in learner language. According to Biber et al. (1999) expressions like *I mean*, when functioning as discourse markers, "typically retain the same interactive function when they occur initially, finally, or medially" (Biber et al. 1999: 1078). In LINDSEI-SW, the frequency of *I mean I* (relative to that of *I mean* in isolation) highlights *I mean*'s function as a frame or utterance launcher, which can be found in contexts of lexical retrieval difficulties:

(43) *(eh) not very no but . she dresses the more . (eh) .. well how do you say (eh) nice I mean I wear baggy clothes and <breathes> skateboard . shoes and everything and she's m= more you know jeans and . sweater* (LINDSEI-SW)

It is difficult to trace the source of the highly significant underuse of *I mean* in LINDSEI-SW, which is also reported for the French-speaking learners in De Cock's (2004) material. As seen in section 4.2.1 above, the frequency numbers for *I mean* makes for an even more significant underuse in the sample from LINDSEI-NO. It seems that *I mean* is not conventionalized in learner language to the extent of native speech, although its fairly high frequency in LINDSEI-SW as compared to other 2-word combinations suggests that it has been conventionalized to a certain extent. It is possible that the use of discourse markers like *I mean* and *you know* is closely related

to general language proficiency, and that the occurrences in learner corpora can be ascribed to only a limited number of highly advanced speakers. This could be investigated through calculating the statistical dispersion of these combinations across the individual interviews (as discussed in section 4.4.1 below), and through an internal assessment of proficiency rather than one based on extra-linguistic criteria (Granger 1998: 9).

4.3.2 Single Clause Constituents, Incomplete Clauses and Phrases

Table 4.15 below shows the recurrent single clause constituents, incomplete clauses and phrases in LOCNEC, divided according to whether or not the combination includes a (complete or incomplete) verb phrase. The verb phrase-combinations need to be considered alongside the full clause-combinations in table 4.9, since they are often part of longer combinations, e.g. *don't know* (n=237), which has a high frequency mainly due to the 3-word combination *I don't know* (n=209), in addition to other and less frequent combinations like *you don't know* (n=11), *they don't know* (n=3) and *I really don't know* (n=3). The same applies to e.g. the combination *like to* (n=97), which is mainly embedded in *I'd like to* (n=62). In the 'miscellaneous combinations'-column, some combinations are sorted according to lexical content, so that combinations like *the end of the* and *a lot of the* are sorted with *the end* and *a lot* rather than the more frequent but lexically emptier *of the*. Through this, it is easier to spot the presence of certain combinations in the corpus and their relations to other, lexically similar, word-combinations, such as e.g. *a lot of the* in relation to *a lot of people*.

LOCNEC:
SINGLE CLAUSE CONSTITUENTS, INCOMPLETE CLAUSES AND PHRASES

Verb phrase

237 don't know
53 don't know I
8 don't know I don't
8 don't know I think
220 to do
174 to go
42 to go to
10 to go and see
167 was a
151 have to
137 which is
133 to be
8 to be able to
117 went to
111 to get
111 to see
106 know I
100 want to
98 was just
97 like to
24 like to go
94 had to
92 go to
86 was really
85 mean I
79 was quite
78 got a
78 had a
77 have a
77 was like
26 was really good
23 wanted to do

Miscellaneous combinations

583 sort of
50 sort of like
44 sort of thing
10 that sort of thing
28 a sort of
26 to sort of
24 that sort of
23 the sort of
23 sort of the
416 in the
24 in the morning
9 o'clock in the morning
24 in the first
9 in the first year
12 in the middle of
7 in the second year
345 and then
310 of the
264 a lot
129 a lot of
14 a lot of people
11 there's a lot of
9 a lot of the
9 quite a lot of
24 a lot more
229 a bit
49 a bit of
19 a bit of a
217 at the
177 as well
174 and the
168 like that
57 things like that
33 and things like that
27 something like that
16 or something like that
17 and stuff like that
153 all the
147 in a
139 lot of
8 an awful lot of
133 on the
132 to the
120 for a
115 kind of
12 that kind of thing
114 it and
102 for the
99 like the
95 of a
95 one of
50 one of the
91 and so
90 the first
89 yeah and
88 and things
35 and things like
85 the time
36 all the time
8 most of the time
81 like a
79 the end
35 the end of

→				
	23	*at the end of*	**9**	**as a foreign language**
	22	*the end of the*	**8**	**English as a foreign**
	7	*the end of it*	**8**	**all over the place**
43	*at the end*		**7**	**in my first year**
78	**at all**		**7**	**teaching English as a**
77	**just like**		**7**	**G C S E**
50	**at the moment**		**7**	**at the same time**
33	**a couple of**			
7	*for a couple of*			
25	**the fact that**			
25	**end of the**			
24	**some of the**			
23	**a little bit**			

Table 4.15: *LOCNEC: Single clause constituents, incomplete clauses and phrases, 2-4 word combinations (freq. >77/23/7)*

4.3.2.1 'Fuzzy boundaries': Formulaic or not?

Table 4.15 underlines the difficulties of defining and detecting formulaic language on the basis of recurrent word-combinations, as discussed in chapter 2, and it is perhaps useful to discuss some of these combinations in relation to formulaicity at this point. Many of the 2-word combinations in the table are verb phrases followed by a *to*-clause, like *like to*, and verbs in the infinitive, like *to do, to go* and *to be*. Others are complete verb phrases like *don't know* and *wanted to do.* These combinations are particularly challenging from a formulaic and comparative perspective, since they

a.) are co-occurrences of two or more consecutive lexical items,

b.) function as one semantic/pragmatic unit,

c.) appear to be prefabricated and conventionalised and

d.) have a frequency of occurrence which is larger than expected on the basis of chance,

and thus appear to meet all the criteria postulated in the working definition of formulaic sequences (cf. section 4.1.2). Still, there seems to be something 'phraseologically uninteresting' (cf. Altenberg 1998) about many of these combinations, which seems to be due to the fact that they, when they are not part of longer combinations, do not convey any semantic or pragmatic information beyond the meaning of the complete or incomplete verb phrase. These combinations seem to be fully compositional, i.e. their meanings are fully interpretable on the basis of the

meaning of their component parts. Non-compositionality was not included as a defining criterion for formulaic language in section 2.2.2, corresponding with Gries' definition of a phraseologism (Gries 2008: 6), and this position should not be modified, as it would exclude the majority of word-combinations considered in this analysis. However, even though combinations like *I think* and *or something like that* may be fully compositional, they are different from the verb phrase-combinations discussed here in that they, as units, also seem to be able to display meanings and functions which somehow go beyond the meanings of the single words in isolation. In practical terms, the overuse or underuse of combinations like *to be* and *wanted to do* are also not likely to be of great consequence for the native- or non-nativelike impression of NNS speech as compared to the NS speech, to the same extent as the more discourse motivated word-combinations. Wray (2002) debates whether the "referential category" of recurrent word-combinations "may be peripheral to the general nature of formulaic sequences" (Wray 2002: 54), and concludes that "at the very least, we may note that the referential function seems in some way different in kind from the socio-interactional and discourse ones" (ibid.: 54-55). As such, these socio-interactional and discourse functions are perhaps of greater significance from the perspective of advanced learner language, since it is likely that quantitative differences and the misuse of word-combinations performing these functions have greater impact on the general impression of the speech of these learners as native- or non-nativelike. It thus seems reasonable to disregard these combinations at this point. Clause fragments and incomplete phrases like *at the*, *in the*, *and the*, *it and, of the, in a* and *of a* are perhaps easier to disregard according to the working definition, as they cannot be said to function as a unit in the text on any (semantic or pragmatic) level. *Of the* was found to be significantly underused by the Swedish learners in section 4.2.1, but since the combination occurs in a wide range of recurrent combinations, e.g. *one of the, some of the, sort of the* and *a lot of the*, this information is not very revealing in a comparative perspective. Other combinations which may be disregarded from consideration at this point, although they give us information on the content of the interviews, are *as a foreign language*, *English as a foreign* and *teaching English as a*, which all take part in the propositional combination *teaching English as a foreign language* (sometimes interrupted by pauses), and *G C S E*, which is an abbreviation incorrectly counted as a 4-word combination. A similar

combination in the LINDSEI material (table 4.16) is the 4-word combination *as an au pair*, which is repeated 8 times. These combinations also seem to meet the criteria for formulaic language, and their frequency support the notion that 'words belong with other words', but they may be considered too context-specific to be of particular interest in a comparative study: "the larger the n-gram, the more idiosyncrasies appear, due to the particular content being described (Barlow 2005: 352).

4.3.2.2 Expressing vagueness and hesitation: sort of, kind of

However, table 4.15 also include many word-combinations likely to serve multiple functions in the text, which are of an interactive nature, and which are thus likely to be contributing to an impression of non-nativelike speech when their patterns of distribution are deviating from the NS norms. Some of these combinations, such as *sort of* and *kind of*, seem to function mainly as vagueness tags (De Cock 2004) in the native speaker corpus, and are often embedded into longer frequent combinations, like *sort of like*, which also function as single clause elements:

(44) *and it's sort of like I mean unfortunately while we were there my friend Belinda twisted her ankle* (LOCNEC)

(45) *but I I was so nervous because <XX> sort of like I I don't enjoy speaking foreign languages to foreigners cos I always think I'm not gonna be as good as they are so I hardly said a word* (LOCNEC)

These are word-combinations which most often convey something beyond their more transparent meaning as modifiers of noun phrases, though these more literal uses can also be found, as seen in (46):

(46) *er who have erm been ill or have some sort of disability and erm . helping them to learn to do things that they used to do before they were ill* (LOCNEC)

The combinations thus take on pragmatic meaning and functions, as seen in examples (44) and (45), where *sort of like* can be said to function as both a planner, in combination with other combinations like *I mean* and the repetitions *I I*, and as a mitigating element for the new information in the following clauses. These pragmatic functions might have appeared as a result of the high frequency of the word-combinations and of the shorter combinations which are embedded in them which will be further discussed below. In addition, since vagueness tags have been reported

to be underused and misused in learner speech (De Cock 2004), these word-combinations are interesting in a comparative perspective, and will be considered further in the following sections, as well as in relation to the single clause constituents, incomplete clauses and phrases in LINDSEI-SW (table 4.16):

LINDSEI-SW:
SINGLE CLAUSE CONSTITUENTS, INCOMPLETE CLAUSES AND PHRASES

Verb phrase

149 don't know
21 don't know if
92 was a
16 was a bit
6 was a good experience
89 have to
81 to do
78 to go
24 to go to
69 think it's
69 want to
65 to see
64 don't think
17 don't think so
63 to get
63 have a
62 was very
58 go to
58 went to
56 had to
56 think it
56 think I
56 think so
22 to do it
17 would like to
6 to get to know
5 know what it is
5 like to talk about
5 to be able to
5 to go to the
5 to look at the
5 to take care of

Miscellaneous combinations

269 in the
15 in the middle
10 in the middle of
230 sort of
184 and then
5 and then they
20 and then I
5 and then in the
150 a lot
94 a lot of
8 have a lot of
5 and a lot of
5 quite a lot of
7 a lot of people
5 a lot of money
137 like that
26 something like that
20 or something like that
21 stuff like that
16 and stuff like
16 and stuff like that
15 like that but
5 or anything like that
6 and things like that
132 of the
18 one of the
119 to be
115 a bit
16 a bit more
107 in a
17 in a way
5 in a way but
102 kind of
97 lot of
95 on the
94 as well
86 at the
86 to the
83 or something
80 all the
78 and the
73 of course
65 and so
6 and so on but
58 the same
57 not really
56 a very
54 about it
54 for a
51 of a
28 a little bit
28 all the time
20 and so on
8 at the same time
8 as an au pair
5 most of the time
5 the rest of the
5 the middle of the

Table 4.16: *LINDSEI-SW: Single clause constituents, incomplete clauses and phrases, 2-4 word combinations (freq. >50/15/5)*

Table 4.16 shows that yet again, many of the combinations from the LOCNEC material appear as recurrent also in the LINDSEI-SW corpus, echoing the similarities from the top 20 lists in section 4.1.1. However, the table also reveals an absence of some of the combinations which did not match on these lists, such as *sort of like* and *a bit of*, which are not retrieved from LINDSEI-SW even with a considerably lower frequency threshold, thus agreeing with De Cock's (2004) reports on the underuse of certain vagueness tags in learner language. The 2-word combination *sort of*, which in itself was found to be significantly underused in table 4.8 above, is not embedded into longer recurrent combinations in LINDSEI-SW that make is pass the set frequency threshold at all, and searches in the learner corpus retrieve only five instances of *sort of like* and two instances of *sort of thing*, which occur 50 and 44 times respectively in LOCNEC. Similarly, *things like that*, which has a frequency of 57 in LOCNEC, only occur 13 times in LINDSEI-SW, and *a bit of* occurs 49 times in LOCNEC and only 9 times in LINDSEI-SW. The 2-word combination *and things*, produced 83 times in LOCNEC and 21 times in LINDSEI-SW, and the 4-word combination *and things like that*, which occurs 33 times in LOCNEC and 5 times in LINDSEI-SW, also add to this picture of underuse.

However, other markers of vagueness seem to be frequently employed in the learner corpus. As could be seen from the frequencies and chi-square results in table 4.8, the Swedish learners overuse the 4-word combination *or something like that (n=20)*, and table 4.16 in this section shows that the 2- and 3-word combinations which are embedded in it, *or something* (n=83) and *something like that* (n=26), are also frequent in the NNS corpus. These word-combinations are likely to serve similar functions to e.g the underused NS combination *(and) things (like that)*, as seen in examples (47) and (48), where *or something like that* and *and things like that* are both used to extend and modify the semantic content of the preceding noun phrase by adding a vagueness tag.

(47) *(eh) I didn't have to take any most students t= you know have to take for example American history or something like that but (mm) . they: told me I could choose what I wanted <overlap /> to study* (LINDSEI-SW)

(48) *yeah <laughs> I'm supposed to read er . the Guardian and things like that* (LOCNEC)

Similarly, the 4-word combination *and stuff like that* (n=16), and its embedded forms *and stuff* (n=41) and *stuff like that* (n=21) are recurrent in LINDSEI-SW, and might account for some of the underuse of the word-combinations including *things* as compared to the native speaker corpus:

(49) *they were well . very open to li= drugs and stuff like that* (LINDSEI-SW)

The Swedish learners also use the word-combination *kind of* more than the native speakers do, with 14.2 occurrences per 10,000 words contrasting with 9.8 occurrences per 10,000 words in LOCNEC. Again, this might be related to influence from American English, where *kind of* is preferred (De Cock 2004: 238), but its frequency cannot be seen to fully account for the highly significant underuse of *sort of* alone. The 3-word combination *kind of thing* is also more frequent in the NS corpus than in the NNS corpus (1.4 per 10,000 vs. 0.4 per 10,000), which shows that this is not a complete compensation for the underused *sort of thing* in LINDSEI-SW.

As discussed above, many word-combinations can be interpreted as 'uncertainty devices' in learner speech, which can partly make up for the underuse of specific vagueness tags like *sort of* and *a bit of* in their strict 'vagueness marking' senses. However, it is also possible that a considerable proportion of the frequency of these frequent word-combinations in learner language may function in different ways from the native-like usage patterns, e.g. as more overt markers of hesitation or lexical retrieval difficulties. Learners might even feel the need to signal their learner status in terms of being vague about propositions made. This assumption, on the other hand, contrasts with the findings of overuse of the "rather forceful" (De Cock 2004: 241) 2-word combination *of course*.

4.3.2.3 of course *and traces of speaker-visibility*

As mentioned in section 4.3.1, the 2-word combination *of course* occurs 73 times in the LINDSEI-SW corpus, while it does not occur at all on the native speaker list, with 33 occurrences in total. The smaller Norwegian sample contains 41 occurrences of *of course*. This indicates a pattern of overuse of this particular combination by the Scandinavian learners, which is also reported by De Cock (2004) to be the case in the French, Japanese, Chinese and Italian LINDSEI (De Cock 2004: 242). De Cock finds that part of this overuse is due to the French learners' misuse of the combination

(*yes/yeah) of course* as a response, often in contexts where the underused *that's right* would have been more appropriate, and where *yes/yeah of course* "may well make learners sound rather over-emphatic and even impolite" (ibid.). However, the Swedish learners do not seem to generally misuse *of course* in such a way, since out of the 73 occurrences only 6 could be found as a first response to a statement or question from the interviewer, as seen in (50) and (51):

(50) *<A> it doesn't make you . you know more careful do <overlap /> you </A>*

<B> <overlap /> <tuts> </B>

<A> think </A>

<B> (em) . <sighs> yes of course it does . </B> (LINDSEI-SW)

(51) *<A> what about your own children is it something that you want them to get involved in painting and drawing </A>*

<B> yeah of course I mean <overlap /> yeah </B> (LINDSEI-SW)

The other uses of *of course* seem to be predominantly found within the introducing frame of a clause somewhere in the middle of a speaker turn (52), or towards the end of a main clause (53):

(52) *<B> (er) and what should be working and . what should be public and what should be private <breathes> and of course the (eh) continual stress of (eh) .. political </B>*

<A> yeah </A>

<B> enemies all around </B> (LINDSEI-SW)

(53) *and (erm) that's (er) sort of (er) in the middle (erm) . in between Edinburgh and (eh) Inverness . so that's (er) in the Highlands .. really (erm) and then we go (eh) visiting . whisky distillery of course* (LINDSEI-SW)

The overuse of *of course* among Swedish and Norwegian learners (*of course* occurs 41 times in the Norwegian sample) is also found in the Swedish and Norwegian components of the International Corpus of Learner English (ICLE), which consists of essays written by Norwegian and Swedish university students of English (Hasselgård 2009: 135). Hasselgård links this overuse to a general tendency for "an interactive writing style with a high degree of writer and reader visibility" (ibid.: 137), and a "high frequency of expressions of modality, opinion and evaluation" (ibid.) in

Swedish and Norwegian learner writing, which has also been noted for this spoken material. This tendency is in writing often linked to a limited register awareness, where learners "overuse features from informal conversation in their written output" (ibid.: 123). Judging from the patterns of overuse which emerge from this study of spoken learner language, such as the overuse of *I think*, *I don't know, I guess* and combinations in which they are embedded, it is possible to argue that (Swedish and Norwegian) learners exhibit a tendency for excessive writer/speaker visibility in their English output in general, which thus also influence their language in spoken conversation. This visibility is also emphasized in House's study of *you know* in EFL, as mentioned above, which concludes that the non-native speakers in her EFL data are primarily "speaker-centered" (House 2009: 183). However, Paquot et al. (forthcoming) find in a comparison of argumentative and academic texts produced by Norwegian learners, that there is a significant decrease in types and tokens which typically mark writer-visibility, e.g. *of course* and *I think*, in the move from the ICLE corpus to the VESPA corpus (Varieties of English for Specific Purposes Database). These findings suggest that such markers are indeed connected to register differences, although it may be argued that because academic writing is a more specialized register than argumentative writing, specific register conventions may be so strong as to override other general tendencies which may still be in use in other written and spoken registers.

It is possible that the high frequency of *of course* occurs as a result of transfer from Norwegian and Swedish, where *jo/ju, naturligvis/naturligtvis, selvsagt/förstås, selvfølgelig/givetvis/visst* are all possible translation equivalents which are likely to be used in conversation[18]. However, since the overuse of *of course* seems to be a tendency for learners regardless of mother tongue backgrounds, this overuse is more likely to be due to a combination of factors, including transfer, a possible higher degree of writer/speaker visibility and markers of subjective stance, and a possible underuse of other words or word-combinations, e.g. other adverbials of stance like

[18] Suggested translations were intuitively selected on the basis of searches in the English-Norwegian Parallel Corpus (ENPC) (http://www.tekstlab.uio.no/cgi-bin/omc/PerlTCE.cgi) and the Google Translate tool (http://translate.google.com/).

obviously. In this way, the overuse of *of course* might be subject to a broader functional repertoire in learner language, which is continually reinforced by its high frequency. Considering the 'forcefulness' of the combination, it is possible that this meaning is not as strongly perceived by learners, and that it is seen as just another way of marking 'modality, opinion and evaluation'. The discourse-internal use of the combination as seen in (52) and (53) indicates that, again, the combination is used in a more inward looking way, functioning as a disclaimer on the learners' his or her own propositions: 'this [previous or preceding proposition] is also an important point to make'.

4.3.3 Repetitions and Filled Pauses

As mentioned in section 4.1.1.1 above, repetitons and filled pauses may perform important functions in spoken discourse, and should thus not be completely disregarded in studies of spoken behaviour. Although the top 20-tables in section 4.1 were not dominated by combinations consisting of filled pauses and repetitions, tables 4.17-4.18 show that repeats and filled pauses are prominent in both the NS and NNS corpora.

LOCNEC: REPETITIONS		FILLED PAUSES	
211	**yeah yeah**	**286**	**erm/er I**
27	*yeah yeah I*	*35*	*erm I don't*
205	**I I**	*26*	*erm I don't know*
126	**it's it's**	**264**	**and er/erm**
118	**the the**	**160**	**but er/erm**
107	**no no**	**113**	**erm and**
106	**and and**		
96	**a a**		
84	**it it**		
28	*it it was*		
47	**was it was**		
44	*it was it was*		
15	**I think I think**		
14	**I don't I don't**		
11	**that was that was**		
10	**I was I was**		
7	**I can't I can't**		
7	**it is it is**		
7	**there was there was**		

Table 4.17: *LOCNEC: Repetitions and filled pauses, 2-4 word combinations (freq. >77/23/7)*

LINDSEI-SW: REPETITIONS		FILLED PAUSES	
182	**I I**	**332**	**eh/er/erm/em I**
21	*I I I*	**251**	**and eh/er**
78	**to to**	*21*	*and eh I*
73	**and and**	**50**	**but eh**
68	**in in**	**38**	**eh/er I think**
59	**they they**	*5*	*er I think I*
58	**the the**	*7*	*and eh I think*
57	**it's it's**	**21**	**eh/erm/em I don't know**
57	**no no**	**18**	**eh I don't**
19	*no no no*	*5*	*and eh I don't*
5	*no no no no*	**17**	**it was eh**
22	**was it was**	**16**	**eh it was**
21	**it was it**	**5**	**I don't know erm**
20	**it was it was**	**5**	**and em then she**
19	**and I I**		
8	**in the in the**		
7	**I'm not I'm not**		
6	**it's a it's a**		
6	**I I don't think**		
6	**I think I think**		
6	**they were they were**		
5	**I don't I don't**		
5	**that was that was**		
5	**in a in a**		

Table 4.18: *LINDSEI-SW: Repetitions and filled pauses, 2-4 word combinations (freq. >50/15/5)*

As it is beyond the scope of this thesis, and since the incomplete data selection does not allow it, no conclusions can be made as to whether there are more repeats and/or filled pauses in native speech or vice versa on the basis of tables 4.17 and 4.18. Considering filled pauses, the data would also have been different had all the different ways of transcribing pauses been conflated, as mentioned in section 4.1.1.1. However, tables 4.17 and 4.18 are interesting in terms of recurrent word-combinations and formulaic sequences in several respects. First, both tables show that the conjunctions *and* and *but* often precede a filled pause. De Cock (1998) describes these combinations as "neglected formulae", which perform a variety of functions (De Cock 1998: 69), and thus emphasize how a corpus-driven method can bring attention to patterns and functions which are difficult to discover on the basis of intuition. Secondly, the tables pick up on some of the words and word-combinations that are highly frequent overall, e.g. *I think, I don't know, it was*, and indicate that these

combinations are closely connected to the possible planning functions and the general hesitation we associate with repetitions and filled pauses[19]:

(54) *<A> so you'd like to go back there <\A>*

<B> erm ... I don't know . I think I think I think yeah <X> .. I probably would like to go back there but I might like to see some of the other . parts </B> (LOCNEC)

(55) *and it was it was (em) . somehow it was easier to: . (er) . not look at everything in black in whi= and black and <overlap /> white* (LINDSEI-SW)

In example (54) from LOCNEC, it seems that both the filled pause, the unfilled pauses, the word-combination *I don't know* and the repeated *I think* are part of the speaker's strategy to keep the words flowing and retain possession of the speaker floor while he or she is planning an answer to the question. Similarly, in the example from LINDSEI, *it was,* although a significant part of the upcoming clause (*somehow it was easier to (...)*), due to the repetition and the filled pause following it, seems to spring to the speaker's mind faster than the rest of the clause, and the whole sequence *and it was it was (em)* thus stalls the utterance until the rest of it is in place. Aijmer (2004) comments on the use of *I don't know* as a 'pause-filler' and 'uncertainty device' in a sample from LINDSEI-SW, and finds that "the uncertainty may be underlined by repetition and by other markers" (Aijmer 2004: 186). Aijmer also find that this use of *I don't know* is predominantly a feature of learner language, although example (54) suggests that this usage may also be found in native speech. It is thus possible that the extremely high frequency of some of the word-combinations in both LINDSEI and LOCNEC is partly due to this 'pause-filler' function, and that the overuse of these combinations in learner language is due to a greater need for such pause-fillers, which in turn comes as a result of greater processing constraints. The high frequency of combinations like *I don't know* in both LOCNEC and LINDSEI-SW (and its reduced form *I dunno* in LINDSEI-SW), which is also increased by such local repetition as seen in tables 4.17 and 4.18, may according to Bybee (2010) lead to "shifts in pragmatic nuances and functions" (Bybee 2010: 44), and it is possible that

[19] Naturally, hesitation is also inherent in the semantics of 'thinking' and 'not knowing', so this might have an effect on the occurrence of repetitions and filled pauses here.

extended pragmatic usage occurs in learner language as a result of higher overall frequencies of the combinations. However, as seen in section 4.3.1.3 above, an extended usage of *I don't know* in its more propositional senses in learner language may also explain the frequency discrepancy.

The fact that these combinations are repeated several times (in both corpora), also indicates a strong connection between these individual words in the speakers' minds, and strengthens the possibility of holistic storage of these sequences in the mental lexicon. However, the overall frequency of the combinations could be seen as both cause and effect of this supposed collocational strength, underlining the fact that frequency cannot be the single determiner for formulaicity. At any rate, tables 4.17 and 4.18 seems to confirm, and extend to include word-combinations as well as words, Biber et al.'s suggestion that "(...) the more frequent a word is, the more readily retrievable it is from a speaker's memory" (Biber et al. 1999: 1059), and that "such a word precedes natural hesitation points in the utterance, and becomes a natural locus for a repeat" (ibid.).

4.3.4 Picture Description Task

Tables 4.19 and 4.20 show the recurrent word-combinations in the material which are strictly related to the part of the interview where the subjects are asked to describe what is going on in a sequence of pictures (see Appendix). The picture descriptions in both corpora vary in size, and the extent to which the interviewer is involved also differs between interviews.

LOCNEC: PICTURE DESCRIPTION TASK	
25	**all her friends**
10	*to all her friends*
10	**she doesn't like it**
8	**a picture of a**
8	**painting a picture of**
7	**and she doesn't like**

Table 4.19: *LOCNEC: Picture description task, 2-4 word combinations (freq. >77/23/7)*

LINDSEI-SW:
PICTURE DESCRIPTION TASK

53	**the picture**	**9**	**shows it to her**
16	*the picture and*	**9**	**sitting in a chair**
5	*at the picture and*	**8**	**she shows it to**
5	*the picture and she*	**8**	**a picture of a**
6	*look at the picture*	**7**	**she wants him to**
52	**she looks**	**6**	**in the first picture**
5	*the way she looks*	**6**	**it doesn't look like**
5	*she looks a bit*	**5**	**off to her friends**
25	**to her friends**	**5**	**doesn't look like her**
9	*to her friends and*		
8	*it to her friends*		
17	**she doesn't like**		
6	*and she doesn't like*		
7	*she doesn't like it*		
5	*she doesn't like the*		

Table 4.20: *LINDSEI-SW: Picture description task, 2-4 word combinations (freq. >50/15/5)*

In example (56) from a picture description by a Swedish learner, the description takes shape of a monologue with no contributions from the interviewer, while the interviewer is more verbally present in example (57) from the native speaker corpus. The recurrent word-combinations appearing as recurrent in tables 4.19 and 4.20 are underlined.

(56) *<B> .. okay .. (er) well . (erm) . in the first picture there is (erm) . there is a man he's probably a painter . (erm) . paints . portraits . (em) and this . lady he paints she looks like an Egyptian . (erm) .. and he's going to paint her and (er) when he's painted her she looks (eh) at the portrait and thinks that (eh) . she doesn't like the hair .. really . and (em) .. she complains about it so he makes another . portrait I I think she she doesn't like (eh) the way she look . she looks (er) . basically but (erm) . so he paints her again and (er) he makes her prettier . and (eh) . she makes her he makes her more .. more beautiful and he makes the the person on the picture smile and (em) . her hair is <starts laughing> more beautiful <stops laughing> and (er) . so (eh) later when when she shows this picture to (eh) to her friends she's very proud and (eh) thinks she looks very beautiful .. yeah <laughs> </B>* (LINDSEI-SW)

(57) *<B> yeah <end_laughter> (erm) the fist er picture is .. of a: an artist painting a portrait of a lady <\B>*

<A> mhm <\A>

<B> and the second one .. the lady er is criticising er the portrait she obviously doesn't like it . and (er) . she's pointing at it and (er) . she's saying what she doesn't like about it .. (er) . this third one (er) .. the[i:] er artist is continuing with his portrait but he's almost finished it .. (er) and .. the[i:] er in the painting er the lady's smiling and in real life she's a: . really miserable face <\B>

<A> <laughs> <\A>

<B> <laughs> and in the final .. (er) picture the[i:] (er) .. the[i:] (er) model is showing off the portrait <u>to all her friends</u> .. and .. and it's a bit the contrast to the second picture where she: looks like she hates the picture <\B>

<A> <laughs> <\A>

<B> and in this one she looks really: proud of it and she's showing it off to everyone <\B> (LOCNEC)

The picture description task is a small part of the corpora, and creates a situation very different from that of the remaining interview. Picture description tasks generally provide many opportunities for comparative research in terms of e.g. narrative structure, and are useful as prompts for the production of certain vocabulary or structures, which is important for studies of e.g. language impairment (cf. Lind et al. 2008). Considering the increased constraints on authenticity this task represents, and how it differs from the rest of the text, this part of the corpora could perhaps have been left out of the present analysis altogether. It is also perhaps to be expected that there are fewer processing constraints involved in this task than in the parts where the speaker has greater choice of conversational content, and where the floor is open for the interviewer to interrupt, and that there would thus be less need here for formulaic language to ease planning and keep the floor. At the same time, tables 4.19-4.20 indicate that task-specific recurrent word-combinations are widespread in this task, particularly in the NNS corpus, where almost the whole story of the pictures can be grasped from reading these combinations only. This is also perceptible in examples (56) and (57), although these samples are too small to draw any conclusions. The examples also show that parts of the language production are not radically different from the rest of the corpus, and that some of the most recurrent word-combinations in the two corpora, e.g. *I think* and *it's a bit,* also appear in this section:

(58) *I <u>I think</u> she she doesn't like (eh) the way she look . she looks* (LINDSEI-SW)

(59) *.. and <u>it's a bit</u> the contrast to the second picture where she: looks like she hates the picture* (LOCNEC)

It thus seems appropriate to include these sections in frequency counts; with the precaution that this might influence the authenticity of the material as a whole.

4.4 Recurrent Word-Combinations: Summary

The previous sections have outlined the most frequent recurrent word-combinations in LINDSEI-SW and LOCNEC, and highlighted major similarities and differences mainly on the basis of frequency information and the structure of the individual word-combinations. The statistical analyses of section 4.2.1 showed that there are patterns of over- and underuse of word-combinations in LINDSEI-SW, and the great type-related similarities of the top 20-lists in section 4.1.1 were confirmed through considering a more extensive data set. The previous sections provided further contextual and co-textual information on the highly frequent combinations, and discussed quantitative and qualitative aspects of some combinations which were not among the most frequent in the corpora. The contextual analysis was prompted by the frequencies and collocational patterns presented, and tentative conclusions were reached regarding the over- and underuse of patterns, such as a greater tendency for speaker-visibility in learner speech, or the influence from American English. Returning to the questions posed after the preliminary n-gram results, it seems that some have been partially answered, while others must be subject to further speculation:

- Why are there so many similarities between the most frequent 2-4-word combinations in LOCNEC, LINDSEI-SW and the LINDSEI-NO sample?
- Are there also noticeable *differences* in the relative frequencies of the highly recurrent word-combinations occurring in the corpora, and if so, what does this entail?
- Are the most frequent word-combinations types typically embedded into other frequent combinations?
- Are there noticeable differences in the functional patterns (usage) of the highly recurrent word-combinations in the corpora, and if so, is this a result of proficiency levels, or other factors?
- Can the highly recurrent word-combinations found in the corpora justifiably be labelled as formulaic sequences, according to the working definition of this thesis (cf. section 4.1.2)?

It seems as if some valid information on the similarities and differences between advanced learner and native English speech can be gleaned from the overall patterns presented above. This shows some of the strengths of the corpus-driven recurrent word-combinations approach; by extracting and presenting large sets of data we can get a good impression of the overall make-up of the register the data is designed to represent. This, in turn, makes it possible to see usage patterns in combination, and debate whether certain explanations can be applied to several findings. However, the approach can also be limiting, in that too much data is considered, restricting the validity of any conclusions reached. This concern is particularly relevant for the last two questions concerning usage-patterns and formulaicity, which evidently require more extensive qualitative analyses. The next chapter will recapitulate and assess some of the assumptions made regarding formulaicity and formulaic sequences in this chapter, and present a more comprehensive analysis of the most significantly overused word-combination in the learner data, *I think*.

4.4.1 A Note on Individual Variation

From looking at the individual interviews and the examples employed in the analysis above, there seems to be considerable variation in the use of word-combinations among the learners, and this variation is not controlled for in this analysis. Aijmer (2009) reports on this in her study of *I don't know* (and *I* dunno) in LINDSEI-SW and LOCNEC, finding that some learners do not use this word-combination at all, while others show an extensive use of it, greatly contributing to its high overall frequency (Aijmer 2009: 155). Prominent individual differences are also found in Aijmer's study of the uses of *well* in LINDSEI-SW (Aijmer 2010: 249), and in House's EFL study of *you know* (House 2009: 180). It also seems that some learners use many of the recurrent word-combinations found in section 4.3 frequently and in combination, as seen in example (60), where the student is talking about a performance put up by his or her theatre group (combinations that have been discussed above are underlined):

(60) *<B> there and and . and (em) . so but the second and the third that's when you <u>sort of like</u> knew more or less <overlap /> how </B>*
<A> <overlap /> (uhu) </A>
<B> what to expect and where the . where the audience would laugh <u>and you know</u> <u>and things like that</u> </B>

<A> yeah </A>

<B> <tuts> and the last few <breathes> I think that's where we sort of got . a bit .. almost I think we'd .. we almost sort of like you know oh well yeah we know how to do this <overlap /> and so </B>

<A> <overlap /> all right </A>

<B> we didn't really pay as much attention as we ou= or try as hard as we perhaps should have </B> (LINDSEI-SW)

The learner in this particular interview reports that he/she 'was a student at the university of Sheffield (em) for four years', which may explain the use of *sort of like*, *a bit* and *you know* in the extract, combinations found to be significantly underused in the corpus as a whole. Similarly, House (2009) reports that in her data "all the speakers who make heavy use of *you know* have spent a considerable time either in an English speaking country or they had ample opportunity speaking ELF at various different stages of their lives" (House 2009: 180). In (58), the frequent use of combinations which are highly recurrent in both the NS and NNS corpora is perhaps a result of the contextualized exposure of spoken English this learner has had the opportunity to access for a long period of time, which is likely to be a contributing factor to the production of nativelike formulaic language in general. In addition, Biber et al. (1999) find that some word-combinations "are associated with personal speech habits, with some individuals making an extremely frequent use of them" (Biber et al. 1999: 1005), even in the language of native speakers, a consideration which is also stressed by De Cock: "one needs to be aware of the fact that formulae can act as idiosyncratic 'lexical teddy bears' (Hasselgren 1994) for some NS and NSS speakers, because it can significantly affect conclusions regarding learner overuse and underuse of formulaic sequences based on simple frequency counts" (De Cock 1998: 75). Electronic dispersion measures, as described in De Cock (1998) have not been calculated for the material in the present thesis, which makes the results vulnerable to idiosyncrasies. However, the fact that combinations like *sort of like* do not appear as frequent in the learner corpus as a whole, despite some learners showing a considerable use of it, suggests that the size of the corpus and the number of interviews are at least to a certain extent large enough to ensure that the quantitative results are representative for Swedish advanced learners of English in general.

5 Formulaic Sequences in Advanced Learner Language

> "(...) identifying formulaic sequences in normal language can be rather like trying to find black cats in a dark room: you know they're there but you just can't pick them out from everything else" (Wray 2009: 101)

The analysis of a large variety of recurrent word-combinations in chapter 4 brings up a number of issues regarding the identification of formulaic sequences. Firstly, the attention to frequency and form brought about by the use of the corpus-driven method inevitably brings attention to many recurrent word-combinations which are not commonly recognized as being conventionalized, holistically stored, or in any way formulaic. Secondly, considering the frequent word-combinations of a corpus without any form of 'formulaic filtering' (De Cock 1998: 71) leads to quantitative and qualitative analyses which also take into account the *non-formulaic* instances of word-combinations which have recognized formulaic status only in certain contexts. However, returning to a usage-based, dynamic framework of language competence, it is possible that considerations of the relation between prototypical and pragmatic uses of frequent word-combinations can provide insight into (i.) the changes of meaning and structural properties the emergence of formulaic language entails, and (ii.) how this process is at work in a language variety which is characterized by very diverging degrees and quality of input and use among its users, such as a foreign language used by learners from a particular linguistic and cultural background.

This chapter will consider the combinations discussed in the previous chapter in terms of usage patterns and formulaicity, and go one step further in trying to explain some of the patterns of over- and underuse found in the data. Since this discussion is of a more qualitative nature, examples from the Norwegian material are considered to a greater extent than in the previous chapter. The next section will however discuss some of the frameworks which can be employed to explain the use (or lack thereof) of formulaic language.

5.1 Motivations and Processes Determining Formulaic Language

The use of formulaic sequences is, according to Wray (2002), "a linguistic solution to a non-linguistic problem" (Wray 2002: 101). This claim serves to shift attention away from the surface form of the sequences, the 'black cats', and rather to the question of

why holistically stored formulaic language is needed and used in language production at all. Wray postulates three overall aims for the formation of linguistic output, "to refer, to manipulate and to access information" (ibid.), and whereas a speaker's aim to be referential will often encourage non-formulaic (novel) output, hearer manipulation and the accessing of information often calls for formulaic language. Wray further proposes that the various functions performed by formulaic sequences can also be collapsed into three: "the reduction of the speaker's processing effort, the manipulation of the hearer (including the hearer's perception of the speaker's identity), and the marking of discourse structure" (ibid.), which seems to correspond well with how many of the recurrent word-combinations discussed in LOCNEC and LINDSEI-SW were found to function. Figure 5.1, reproduced from Wray (2002: 97), shows these three functions in further detail, as well as their relation to the overall concern of benefiting the speaker:

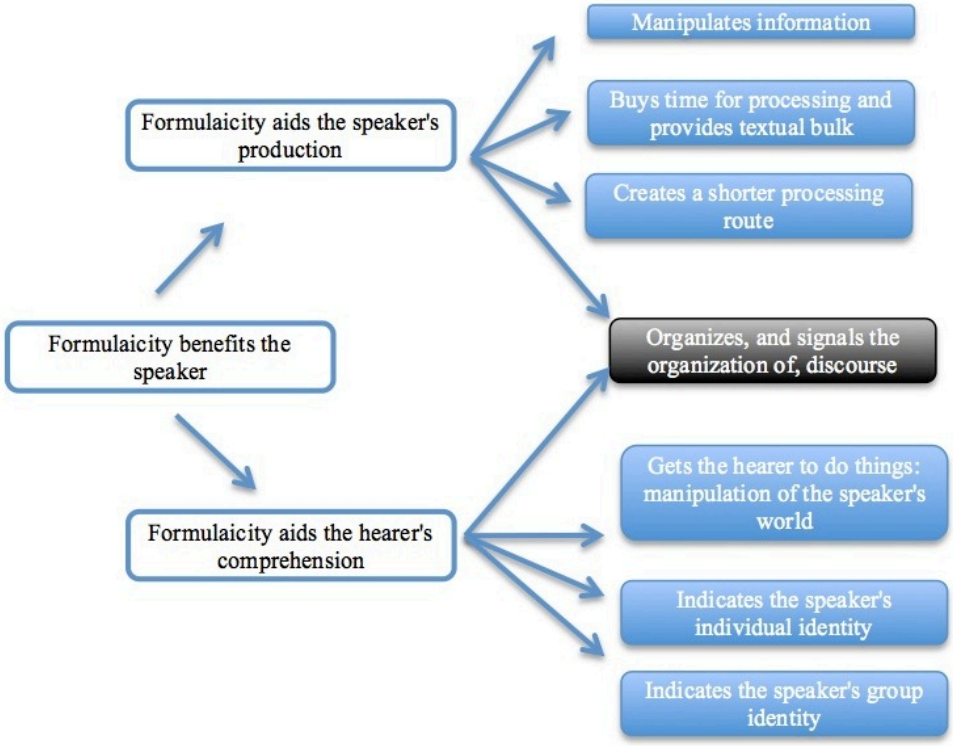


***Figure 5.1**: The functions of formulaic sequences (cf. Wray 2002: 97)*

The linguistic problem-solving is, according to Wray, tackled in different ways, according to the language users' individual assessment of the situation and his or her abilities: "The three dimensions – processing, interaction and discourse marking – all

operate largely independently, and so each person, in each unique situation, will apply slightly different selection criteria to a slightly different set of options, from those available to anyone else" (ibid.). The present study rather attributes tendencies to similar choices across a range of individuals, so that the 'set of options' individual learners possess is considered to be similar across this language population (see section 4.4.1 on variation). In this way, some non-nativelike patterns of distribution, such as the reliance on only a few word-combinations in contexts where native speakers make use of several, can be attributed to the restricted lexical inventory of advanced learners in general. This is perhaps a simplification, but focusing on probabilities still seems to be a useful perspective, particularly from the point of view of language acquisition.

In learner language, attention to linguistic forms is important in a way which is perhaps peculiar to studies of formulaicity in non-native speech, since aspects of form overtly marks native- or non-nativeness. In this perspective, the 'black cats' which are easier to point out on the basis of e.g. their abnormally high frequency, strange behaviour or peculiar shape, hold particular significance since they can be undesirable from a learner's point of view. However, in order to explain the occurrence of formulaicity in learner language output, it is useful to consider Wray's factors considering the learners' needs and motivations, since "insofar as the learner's communicational agenda and processing priorities differ from those of a native speaker, this will create a different set of formulaic sequences, and lead to a different use of them" (Wray 2002: 194). Figure 5.1 seems to cover most of the explanations discussed in terms of the quantitative and qualitative findings above, perhaps most notably the ones relating to aid of speaker production. The 'manipulation of the hearer'-boxes can be said to include the discussions on speaker-visibility (section 4.3.2.3) and vagueness/hesitation marking, since learners (consciously or unconsciously) can be said to use word-combinations like *I think, I guess, I don't know* and *and stuff like that* to (i.) reduce his or her own responsibility towards propositions or claims made and (ii.) to signal his or her learner status and ensuing lexical retrieval difficulties to the hearer.

Individual motivations, needs and limitations may thus serve as explanations for the production of formulaic sequences overall. However, they do not seem to fully

explain why formulaic language use occurs and changes with exposure to the target language. Word-combinations which often re-occur, and which are subject to holistic storage, may often lose their compositional nature and act in ways which separate them from the meaning of their component parts. In addition, some formulaic sequences are fully compositional, but have taken on a set of extended meanings and functions, which has been found to be the case for combinations such as *I don't know* and *I think* in spoken language. Bybee (2010) refers to domain-general processes in our cognition to explain the changes of meaning and function which sometimes occur with words and word-combinations in language use. These processes are independent of language and also perform other cognitive tasks, but are believed to influence language output and encourage language change. Some important processes for the emergence and change of formulaic language are *chunking, categorization, rich memory storage* and *cross-modal associations* (Bybee 2010: 7-8). Chunking, "the process by which sequences of units that are used together cohere to form more complex units" (ibid.: 7), most often occurs through repetition, and is evidently at work in formulaic language processing. When formulaic sequences are used or encountered, contextual experience is stored in the rich memory storage, and mapped onto linguistic forms by means of categorization (ibid.). In this way, formulaic sequences are stored alongside its phonetic, semantic and contextual information, and are constantly entrenched and renewed as new experiences with them are similarly categorized: "each experience with language has an impact on cognitive representation" (ibid.: 7-8). Cross-modal associations, in turn, make possible the associations between linguistic output and co-occurring events, so that "inferences made from the context of particular utterances can (...) come to be associated with particular sequences, giving rise to changes in meaning" (ibid.: 8). These domain-general processes may thus help explain the occurrence of formulaic sequences, but also their dynamic nature, constantly revised during language processing: "change occurs gradually" (ibid.: 114).

Since these are language-independent processes, and since it was assumed in section 2.2.3 above that both native and learner language are subject to the same possibilities and limitations, it should be assumed that the processes of *chunking*, *categorization*, *rich memory storage* and *cross-modal associations* are also at work in the processing

of a non-native language. This aspect is stressed in Bybee (2009), where native-like exposure to the target language is underlined as the primary concern in foreign language learning and teaching, although other factors, like the motivation and ability for learning, are also important (Bybee 2009: 232). Since learners are most often exposed to a limited and diverging input and competition from forms and categories in the mother tongue, it is also likely that processes of chunking and what is represented in the rich memory storage in a learner language population can diverge from those present in a native language population: "Because each instance of language use impacts representation, variation and gradience have a direct representation in the language-user's system" (ibid.: 9). This could thus lead to differences in the usage patterns of frequent formulaic sequences, which has been assumed in relation to the highly frequent word-combinations found in LOCNEC and LINDSEI-SW. Aijmer's (1997) definition of pragmaticalization of word-combinations (section 4.1.1.1) can be said to related to the early stages of grammaticalization, as described by Hopper and Traugott (2003):

> "The potential for grammaticalization lies in speakers attempting to be maximally informative, depending on the needs of the situation. Negotiating meaning may involve innovation, specifically, pragmatic, semantic, and ultimately grammatical enrichment" (Hopper and Traugott 2003: 98).

This 'pragmatic innovation' is thus related to the processes of change which continually affects meaning and possible functions of words and word-combinations. The idea that usage patterns in learner language need not necessarily be taught or inspired by the target language, and that grammaticalization/pragmaticalization processes can operate at different levels and at different speed, is a concern which is highlighted by both Wray and Bybee. Although Bybee mainly considers abstract, discontinuous 'constructions', it seems plausible that these processes can also be at work for continuous lexical word-combinations, particularly highly frequent combinations, as the ones studied here, since "association by contiguitiy allows forms to take on meaning and allows meaning to change from association with context and with frequently made inferences" (Bybee 2009: 221).

5.2 Traces of Formulaicity in LINDSEI-SW and LOCNEC

In section 4.3.2 on single clause constituents, incomplete clauses and phrases above, some recurrent word-combinations found in the NS and NNS corpora were discarded as candidates for formulaic sequences, as they were judged to be of limited interest from either a formulaic (*of the, go to*) or a contrastive perspective[20] (*as an au pair*), or both. Some combinations were described as being 'formulaically uninteresting', on the basis of a further narrowing down of the initial, broad definition of formulaic sequences. Consequently, combinations which did not seem to display any meaning beyond their 'literal' or propositional meanings were disregarded, along with combinations which were fragmentary and hence difficult to identify as units. Lists of frequent word-combinations as the ones presented in tables 4.9-4.10 and 4.15-4.20, as well as the above discussion, address the question of whether frequency is a defining feature of formulaicity, and whether frequent word-combinations like *of the, for a* and *to the* are holistically stored in our mental lexica, even though they do not seem display much unified semantic or pragmatic meaning of their own accord. For practical purposes, and for the purpose of the comparative analysis, it seems reasonable to refrain from further comment on these combinations, but including them in a frequency search as the one presented above does illustrate "the pervasive and varied character of conventionalized language in spoken discourse" (Altenberg 1998: 120), where "the use of routinized and more or less prefabricated expressions is evident at all levels of linguistic organization and affects all kinds of structures, from entire utterances operating at discourse level to smaller units acting as single words and phrases" (ibid.).

It is possible that data such as that presented in chapter 4 "emphasizes rather than clarifies the fuzzy character of phraseology" (ibid.: 121), but it is at the same time useful to test phraseological definitions against such a broad selection of data. Considering the very 'fuzzy' boundaries of the definition, such as the strings with filled pauses in section 4.3.3, if only briefly, might benefit our understanding of formulaicity and language processing in general, and provide suggestions for further

[20] I.e. the comparison of the NS and NNS corpora.

research. As we gain more knowledge on prefabrication of language, smaller and fragmentary words and word-combinations which are difficult to assign to meaning or function may also be considered to be formulaic, with all the implications that might entail for our notion of linguistic competence: "indeed, any string of words might turn out to be formulaic" (Wray 2009: 11). This possible pervasiveness of formulaicity has implications for how we view language processing and language acquisition - and is likely to influence e.g. pedagogical ideas about language learning and teaching. While accepting that formulaicity is likely to extend well beyond that which is attributed to the word-combinations discussed here, the scope of the following discussion is narrowed considerably, and only combinations which in certain contexts can be attributed to one or more interpersonal or textual function(s) are taken into account. Such a unified function is thus, along with a frequency which is greater than what would be expected on the basis of chance, a plausible indication of holistic storage and formulaic status, in a scope which is perhaps as far reaching as possible when studying linguistic output only. This is also the stance taken by De Cock et al. (1998), who confine their study to "frequently used multi-word units that perform pragmatic or discourse structuring functions" (De Cock et al. 1998: 67).

5.2.1 Overuse, Underuse and Formulaicity

A summary of the overuse and underuse of word-combinations discussed in the quantitative and contextual analyses of chapter 4 is listed in table 5.1 below. Words in brackets represent some of the most common overused or underused collocational patterns of these word-combinations, in which smaller combinations are embedded.

Overuse	**Underuse**
I think (it's/so)	*I mean (I)*
I don't (know/think)/I dunno	*you know*
I guess (so)	*that's right*
(or) something (like that)	*(that) sort of (like/thing)*
(and) stuff (like that)	*(a) bit of*
of course	*(and) things (like that)*
kind of	

Table 5.1: *Over- and underused word-combinations in LINDSEI-SW as compared to LOCNEC*

The collocational variations of the combinations in table 5.1 indicate a flexibility of recurrent word-combinations which challenges a strict structural definition of formulaic sequences predominantly based on form. Altenberg finds in his study that "there are comparatively few examples that are completely 'frozen', semantically or grammatically" (Altenberg 1998: 121), and this seems to be confirmed by tables 4.9 and 4.8 in terms of e.g. lexical expansion (*I think* and *I think it's*) and verb conjugations (*I don't* and *I didn't*, *I want to (do)* and *I wanted to (do)*). Less frequent expansions are not easily picked up in a study such as the present one, and discontinuous combinations (cf. *I really don't know*) are also disregarded, mainly due to the n-gram method used. Collocational patterns and the embedding of short combinations into larger ones can tell us something about the distribution and position of combinations in context, and determine whether shorter combinations are frequent mainly as a result of embedding into longer combinations. These observations thus question whether e.g. highly frequent 2-word combinations like *I don't* are formulaic, since their high frequency of occurrence is mainly due to embedding into longer combinations of formulaic nature, such as *I don't know* and *I don't think.* The appearance of *I don't* as one of the combinations which are most often repeated in both LOCNEC and LINDSEI-SW (cf. tables 4.17 and 4.18), is one of the findings which support the idea of *I don't* as a holistically stored combination which is considered a single unit by both native and non-native speakers of English. The examples of *I don't* in initial position and surrounded by discourse markers or other markers of hesitation (61-62), suggests that it is used as an utterance launcher as well as expressing personal reference and negation.

(61) *oh . well I don't I don't do that* (LINDSEI-SW)

(62) *erm no .. erm I don't I don't I don't really get the chance to see her* (LOCNEC)

Following this assumption, other frequent combinations which often occur in initial position in a clause, such as *I think it's* and *I think that's*, are also possible candidates for holistic storage.

In a contrastive analysis, collocational information can also be used to discover potential differences in the formulaic status of certain word-combinations in the two language populations, in this case native- and non-native speakers. It could be seen

above that the 2-word combination *kind of* is overused by the Swedish learners, whereas the longer *(that) kind of thing* was found more often in the native speaker data. This may indicate that the latter combination is not holistically stored in the Swedish learners' memory systems, even though the former is overused and most likely left without internal analysis.

5.2.1.1 Formulaic 'filtering'

The combinations listed in table 5.1 are considered to be good candidates for formulaic sequences in either learner or native English speech, or both, on account of their high frequencies and the fact that they have previously and in this material been found to be used for interactive or discourse organizational purposes. In addition, their diverging frequencies suggest non-nativelike behaviour on the part of the Swedish learners. According to the narrowing of scope in the beginning of this section, some of the specific tokens of these sequences are not formulaic, as they do not 'perform pragmatic or discourse structuring functions'. This include propositional uses of *you know, I don't know* and *I think,* although, as seen above, this distinction can be difficult to determine. De Cock et al. (1998) disregard the tokens of recurrent word-combinations which do not serve any pragmatic or discourse functions, such as the purely referential *you know* (De Cock et al. 1998: 75), and Aijmer (2009) disregards several instances of *I don't know/I dunno* in her study for similar reasons, as her study is strictly related to 'pragmatic markers'. In De Cock (1998) it is argued that "a comparison based on unrefined frequencies may paint a disorted picture of the use of formulaic expressions by NSs and NNSs" (De Cock 1998: 73). De Cock finds, after manually 'filtering' a list of recurrent word-combinations extracted from samples of LINDSEI-FR and LOCNEC, that several instances of some of the most frequent word-combinations do take on literal meanings (ibid.). In the context of this study, this difference can be illustrated by the 2-word combination *kind of*, which was found to be overused by the Swedish learners in section 4.2. In De Cock's material, a highly significant proportion of the occurrences of *kind of* produced by the French learners was judged to be non-formulaic (x^2=11.34). Since the overuse of *kind of* is in the present study assumed to be part of the explanation for the significant underuse of *sort of* (see section 4.3.2.2 above), it is perhaps fruitful to see these combinations in

relation. Examples 63-65 show *kind of* and *sort of* in what can be considered a scale from literal to formulaic use, in the sense described by De Cock (2004):

(63) *so that's I enjoy it most of the time because most most of the times . people are very friendly . (er) and they (eh) . they (er) really enjoy being on a trip . and on such a trip a coach trip because the they they have they know they have bought this kind of trip and they know that . (erm) . it's it's a special . type of organisation wh= when it's such a . like a coach party* (LINDSEI-SW)

erm and .. it . it shows how erm . a a group of boys inspired by a sort of . unauth= . unauthentic sort of English t not unauthentic but erm . (LOCNEC)

(64) *(erm) and he (mm) stands kind of stands back (eh) his <swallows> fist on his hip (er) . I don't think he's too pleased with her criticism* (LINDSEI-SW)

who wants to be king .. and so basically he sort of plots against the king . and he sort of .. plots with the hyenas who are sort of the bad guys in the story (LOCNEC)

(65) *(er) so . (eh) well Boulder is . (em) I don't know how many people live in Boulder but <breathes> (erm) it's more a . (eh) it's a very laid-back place and (eh) . (em) . <tuts> well what can you say it's (em) . (eh) close to the nature kind of* (LINDSEI-SW)

so I'd like to sort of do a bit of<'> exploring sort of (LOCNEC)

In (63), *kind of* and *sort of* seem to be used in their most literal senses, modifying nouns in a way which is compatible with the dictionary listings of *sort* and *kind.* In these examples, the combinations are also preceded by a demonstrative pronoun and an indefinite article, which underlines their attachment to the noun. In (64), however, the combinations are modifying a verb, and seem to function primarily as markers of vagueness or "imprecision" (Biber et al. 1999: 871), or as devices to maintain fluency within the speaker turn. In (65) *kind of* and *sort of* are left out of the main utterance altogether, and tagged on at the end, making for an even more structurally and functionally flexible use of the word-combinations. Leaving *sort of* and *kind of* out of the utterances in (63) would also change the propositional content of the utterances in a more fundamental way than in the two remaining examples. As a further illustration of the discourse structuring use of these combinations, the below examples show *kind of* and *sort of* used as particles marking quoted speech or thought:

(66) *but we just kind of . no we can't talk about this (LINDSEI-SW)*

<B> so er . but then when they left I was sort of it really sort of sunk down <X> I'm here and I'm here for another eight months and I'm on my own and like when you go out the door and everybody speaks French and it's not your language so <\B>

<A> uhu <\A>

<B> you're sort of oh oh no <\B>

<A> <laughs> <\A>

<B> I don't dare speak to anybody <\B> (LOCNEC)

The literal uses of *sort of* and *kind of,* as seen in (63), should thus perhaps be left out of an analysis of formulaic sequences. In some instances, it can also be more difficult to determine the formulaic status of a word-combination, like in (67), where it is not certain whether *sort of* (despite the lack of inflection) refers to different types of positions, or whether it is a structurally flexible insert marking vagueness, which is a more likely interpretation of the remaining instances of *sort of* in the utterance:

(67) *as an assistant erm there are four sort of positions that you can apply for in the department and it's sort of open to everybody so you: you've got so much chance than anybody else . erm and you sort of apply before March and then just wait and see so I'm gonna apply for that and then hopefully and then I've I've applied for working in this English institute in France in in Strasbourg* (LOCNEC)[21]

By considering and interpreting all instances of these word-combinations manually, it would be possible to make firmer claims about whether frequency patterns are mainly due to more literal use of the word-combinations, or to interactive and discourse-related functions. However, automatically extracted collocational patterns can provide at least some indication of usage, and uncover the presence of embedding, such as the frequently occurring *sort of like* and *sort of thing* in LOCNEC (cf. table 4.15). This was also seen in the discussion on *I think that* in section 4.3.1.2. In addition, considering the processes of pragmaticalization/grammaticalization described in Aijmer (1997) (cf. section 4.3.1.2 above) and Bybee (2010), it seems that formulaic

[21] Examples 63-67 also show the general pervasiveness of sort of in British English speech, as the combination is repeated more than once in all four examples.

and non-formulaic usage of word-combinations can be closely related, and it thus seems useful to also consider the more prototypical occurrences as influencing meaning and use in both native and learner language. Quantitative and qualitative differences between learner and native speaker use of the word-combinations which can potentially function formulaically in the sense discussed here, can perhaps tell us something about the processes of pragmaticalization which could not be uncovered otherwise.

5.2.2 Possible Explanations for Quantitative and Qualitative Differences

It was suggested in the discussion in 5.1 that the reasons for the overuse and underuse of particular word-combinations in learner language may be attributable to a number of factors, also related to form, but that considerations of internal and external nature might tie some of the emerging patterns together. Interactional and discourse-organizational motivations for the production of formulaic sequences, as well as cognitive processes creating, facilitating and changing formulaic language, are useful considerations in the task of mapping out the meanings and functions of particular sequences. Suggested explanations for the overused and underused combinations in table 5.1, as well as assumptions about diverging quantitative and qualitative findings, may be summarized in the following way (partly based on the explanations listed in figure 5.1 on why formulaic sequences are used (socially) and how they work (cognitively)):

1. The reduction of the speaker's processing effort (buys time for processing and provides textual bulk, creates a shorter processing route): Greater planning difficulties cause learners to make use of certain highly frequent word-combinations as 'pause-fillers' to ease fluency and lexical retrieval (e.g. *I don't know, I think*). These planning difficulties also lead to a greater use of 'utterance-launchers' or 'frames' (e.g. *I think it's*).
2. The manipulation of the hearer (indicates the speaker's individual identity, reduces speaker responsibility): There is generally a greater tendency for 'speaker-visibility' in learner speech (while native speech is more oriented towards the hearer), and this tendency is connected to a high frequency of expressions of modality, opinion and evaluation in learner language (e.g. *I*

think, I don't know, of course). The preference for these expressions suggests a lack of other ways to introduce arguments and opinions, or a general preference for personal reference among learners of English. In addition, there seems to be a fear of being too assertive about opinions and propositions made among learners, which is related to the learner situation and a generally lower confidence caused by greater processing difficulties. This, in turn, leads to an extended use of different markers of vagueness (e.g. *I guess, or something like that, and stuff*).

3. Lack of pragmaticalization: A high-frequency of certain word-combinations is often linked to unified meaning and pragmatic change, and a limited exposure to contextualized native language use in general and formulaic language use in particular might delay this development in learner language and thus lead to underuse (e.g. *I mean, you know*).
4. Excessive pragmaticalization + restricted lexical inventory: Since learner language consists of a more restricted lexical inventory and a dependency on certain 'islands of reliability', these particular 'islands' are highly frequent and may be subject to similar or increased pragmaticalization processes compared to those of native language development (e.g. *I don't know, I think*). This widening of functional scope of certain sequences also leads to the underuse of sequences which perform similar functions in native language (e.g. *that's right, I mean*).
5. Simultaneous formulaic and non-formulaic treatment of word-combinations: Learners are at a different stage in the pragmaticalization process of highly frequent word-combinations because of their restricted input and application, which leads to a greater tendency for prototypical usage of these combinations, in addition to the more formulaic and pragmatic meanings, which in turn leads to high frequencies (e.g. *I don't know, I think*).
6. Diverging input: Influence from American English input causes diverging usage patterns compared to the speech of a British English native speaker population, as represented in LOCNEC (e.g. *kind of*, *I guess, and stuff*).

'Learners' in this context refers to Swedish and Norwegian learners, and some findings and explanations might in some way be restricted to or emphasized in those

learner populations as compared to other learners, such as the presumably extensive influence from American English. Some of the explanations are slightly contradicting, most notably numbers 3, 4 and 5, which represent the diverging results of this study and others. It is difficult to say why some combinations should be subject to pragmaticalization processes and broad usage pattern while others are not, but a combination of factors of both 'internal' and 'external' nature might serve to explain these diverging patterns.

The second explanation is related to studies of written English learner essays from Swedish and Norwegian speakers (cf. section 4.3.2.3), and on the fact that many of the overused word-combinations displaying speaker-visibility and expressing modality, opinion and evaluation in learner writing are also found to be overused in speech: "(...) they [learner writers] overuse subjective interpersonal metaphors which contain first person references and references to the writer's mental processes" (Herriman and Aronsson 2009: 113). The first person pronoun is in itself more frequent in LINDSEI-SW as compared to LOCNEC, with 533.8 occurrences per 10,000 LOCNEC compared to 571.7 per 10,000 in LINDSEI-SW, a tendency also reported for Swedish learner writing (Ringbom 1998: 46). Herriman and Aronsson (2009) argue that Swedish learners use formulaic expressions to "compensate for the NNS' lack of sufficient knowledge in the foreign language" (ibid.: 116), particularly "knowledge of textual organization in English" (ibid.: 117). It is possible that this inclination to express first person reference and the lack of knowledge about alternative structures and word-combinations is perceivable also in the organization of spoken language, and that these factors are contributing to the overuse of a smaller number of sequences.

To test some of these explanations, the next section will consider one of the word-combinations from table 5.1 in greater detail. The quantitative and contextual features of *I think* in both LINDSEI-SW and LOCNEC were discussed in section 4.3.1.2 above, and some of its meanings and functions were considered based on Aijmer's (1997) and Kärkkainen's (2003) studies of *I think* operating as a partial or fully-fledged discourse marker. It was suggested that the extremely high frequency of *I think* in LINDSEI-SW is due to a combination of explanations 1, 2, 4 and 5 above. The overuse of the longer word-combinations *I think it's* and *I think that/that's* was

further seen to be a possible result of the greater need for utterance launchers and a greater tendency to modify propositional content in learner language. In addition, it was suggested that *I think* is used in a variety of contexts in learner language due to a lack of other ways to express modality and doubt, and that this might in turn lead to a broader range of pragmatic functions. These assumptions call for a closer look at the occurrences of *I think* in context, and a qualitative analysis will perhaps provide some answers which are valid also for the functions of some of the other sequences in table 5.1.

5.3 Qualitative CIA of Formulaic Sequences: I think

It seems clear from the quantitative results above that *I think* can function as a holistically stored unit in native and learner language. Furthermore, it is possible that longer combinations where *I think* is embedded, such as *I think it's*, are also stored holistically, due to their high frequencies. The prototypical and epistemic uses of *I think* may also be considered to be formulaic, but as seen above, it is easier to justify a classification of *I think* as a formulaic sequence in the instances where it clearly functions as a unit, and takes on discourse-oriented functions. However, the fact that literal uses co-occur with more prominent interactive and discourse-structural uses indicate that *I think* is part of an ongoing process of pragmaticalization in both learner and native speech. Judging by the fact that *I think* is significantly overused by both Swedish and Norwegian learners of English, it is possible that this combination, and its expanded forms, might be at a different stage of pragmaticalization in NNS speech than in native language. The initial hypotheses of this section are thus (i.) that the frequency of *I think* reflects the general tendency for speaker-visibility and personal expression of opinion in learner language, and (ii.), that *I think* and the combinations in which *I think* is embedded are used as 'islands of reliability', since non-native proficiency levels lead to lack of other ways to start off utterances or to express epistemic stance and the weakening of speaker responsibility. Thirdly, these two factors may lead to a strengthening of the pragmaticalization process of these combinations, which in turn leads to an even higher frequency, and to an extended set of possible functions, positions in the clause, and collocational patterns. These functions also include the use of *I think* and its extended forms as 'pause-fillers'.

5.3.1 Discourse-functional and Interactive Properties of I think

As mentioned, related to the discussion of pragmaticalization of *I think* is its possible function as a discourse marker. If *I think* can be classified as a discourse marker, this should also have implications for the possible positions this combination can appear in in spoken discourse. Kärkkainen (2003) discusses whether *I think* functions as a discourse marker in native American English, according to a range of linguistic conditions provided by Sciffrin (1987):

i. it has to be syntactically detachable from a sentence; ii. it has to be commonly used in initial position of an utterance; iii. it has to have a range of prosodic contours e.g. tonic stress and followed by a pause, phonological reduction; iv. it has to be able to operate at both local and global levels of discourse, and on different planes of discourse; v. this means that it either has to have no meaning, a vague meaning, or to be reflexive (of the language, of the speaker)

Table 5.2: *Linguistic conditions allowing for a word-combination to be used as a discourse marker (reproduced from Schiffrin 1987: 328; cited in Kärkkainen 2003: 175)*

These conditions are thus contrasted with the more literal uses of *I think*, as either referring to cogitation or opinions and beliefs. However, even if *I think* can be seen to fulfil condition (v.) on meaning, it is still not completely void of meaning: "there is *some* semantic meaning to *I think*, as speakers do not simply choose *you know* or some other discourse marker in its place - in other words we cannot say that *I think* has no meaning at all." (Kärkkainen 2003: 177). In this way, *I think* may still be compositional and evoke the separate semantic meanings of its component words in the speaker's mind.

Regarding the functions of *I think,* Aijmer (1997) states that "position [in the utterance] seems to be important" (Aijmer 1997: 24), which echoes condition (i.) and (ii.) in table 5.2 and, to a certain extent, condition (iv.). If *I think* operates 'on a global level', signifying e.g. planning difficulties, or marking general vagueness relevant to the clause as a whole, it should be able to operate freely in different

positions in the clause. Chapter 4 showed some of the most common collocations with *I think*, and the high frequency in all three corpora suggested that the combination is indeed flexible and versatile as far as context and position is concerned. For ease of reference, the collocational patterns are presented once more in table 5.3:

LINDSEI-SW	LOCNEC
I think	***I think***
(no) I think I (would)	*I think I*
(and/so/but) I think it's (a/more)	*I think it's (a)*
I think it (was)	*I think it (was)*
(and) I think that's (the)	*but I think*
(yeah/yes) I think so	*yeah I think*
I think that (was)	*and I think*
I think the	*I think I think*
I think they	
I think and	
but I think	
yeah I think	
and I think	
so I think	
no I think	
I think I think	
(and) eh/er I think (I)	

Table 5.3: *The recurrent 2-4 word combinations with <u>I think</u> in LINDSEI-SW and LOCNEC (freq. >50/15/5 & >77/23/7)*

The collocational patterns with conjunction or response items (*but/yeah/and/so/*no), particularly in LINDSEI-SW, go a long way in suggesting that *I think* often occurs at the beginning of 'thought units', perhaps more so in NNS speech than in NS speech. In addition, the recurrence of *I think and* in the NNS column suggests that *I think* can also occur at the end of a clause, before the framing stage of a new thought unit. However, the collocational patterns do not provide enough information on the position of *I think* in the clause. In epistemic use, the contextual information is important, e.g. *I think* collocating with *that/that's/it's,* but these combinations might also have interactive or discourse-organizational functions, which are difficult to determine from collocational information only. It is also possible to assume from table 5.2 that since a fully-fledged discourse marker can potentially appear anywhere in the clause, being 'syntactically detachable from a sentence', this is part of the reason why *I think* does not appear in many conventionalized collocational patterns in

LOCNEC. There thus seems to be a need to consider whether *I think* typically displays the kind of structural diversity associated with discourse markers. Aijmer (2009) claims in her investigation of *I don't know* in LINDSEI-SW and LOCNEC that "the position of *I don't know* in the clause is pragmatically or interactively motivated rather than syntactically and is therefore an important cue to its function" (Aijmer 2009: 154). As mentioned in section 5.2.1.1, Aijmer has discarded from her study the cases where *I don't know* takes its literal meaning (e.g. *I don't know if (...), I don't know the word*). Since a similar filtering has not been conducted for the analysis of *I think* in this study, some of the uses of *I think* might indeed be more syntactically motivated, but it seems fruitful also to include these instances, as it seems that some instances might appear as a result of both interactive and content-related motivations simultaneously.

It is also possible to assume that if there is hesitation surrounding *I think*, or if the combination is interrupted or the clause is started over, the combination is used as a 'first thing that sprung to mind'-option, filling pauses and launching the utterance while signalling that propositional content is about to appear. The repeated *I think* in both corpora, and the filled pauses which often co-occur with *I think* in the NNS corpus, as seen in table 5.3 are indications of this function.

5.3.2 Analysis

The following analysis is primarily of a qualitative nature, although references are made to the frequency findings of chapter 4, and percentages are calculated but not statistically compared, primarily due to the small data set considered. To limit the material, I extracted one third of all the instances of *I think* in the three corpora, and the resulting data set thus contains 144 instances from LOCNEC, 177 from LINDSEI-SW and 58 from LINDSEI-NO. The combinations were extracted in running order, and thus belong to a limited number of interviews only, but are nonetheless believed to be to a certain extent representative of the different speaker populations. The instances were then classified according to position in the clause, and notes were made as to whether clauses starting with *I think* were truncated, or whether there were any hesitation markers surrounding the combination.

Since instances of *I think* were counted according to position in the clause rather than position in the utterance, examples such as the one seen in (68), where the combination occurs in the middle of an utterance and after a co-ordinating conjunction, is counted as 'front':

(68) *(...) I feel like the Danish people they're . they'e enjoying life . and (eh) I think some <laughs> I thin= I think some (em) .. (eh) studies have been made (...)* (LINDSEI-NO)

Since some utterances extend across several lines in the transcription, it would for this purpose be difficult to determine the position of *I think* in terms of the higher-level 'utterance'. For similar reasons, no distinction was made between independent and subordinate/co-ordinate clauses. The language of spoken conversation is perhaps best described in terms of clausal units, as described in Biber et al. (1999: 1069), operating within a complex system of embedding and coordination. Altenberg (1990; 1998) similarly regards speech in terms of a linear composition rather than a hierarchical one. For this purpose, position is loosely determined on the basis of perceived 'thought units (see section 4.1.1.1), where the beginning of a thought unit marks a potential for greater planning difficulties, and combinations occurring at the end of one typically serves as a comment on the thought unit as a whole. In syntactic terms, *I think* positioned either in front of obligatory clause elements like subject and verb are present in the 'front' column. Mid-position refers to the times where *I think* is inserted within a phrase or between two clause elements:

(69) *suddenly you were in charge of thirty (eh) I think twelve years= year olds seventh graders* (LINDSEI-NO)

Inserted clauses beginning with *I think* are listed separately (70-71), and so are cases where *I think* takes part in nominal (72) or postmodifying relative clauses (73):

(70) *I remember all the: I don't know the: English word but the: adult m= (eh) wind band I think it's wind band or wind ensemble or something like that* (LINDSEI-NO)

(71) *and I was it the first time I think it was in ninety-one . April ninety-one . and: (...)* (LINDSEI-SW)

(72) *So I explain what I think is going on* (LOCNEC)

(73) *Yeah eh the film . that I think is particularly good is Dirty Dancing* (LOCNEC)

The latter category thus illustrates a particular use of *I think* where the combination most likely does *not* function as a discourse marker.

Sometimes it is difficult to determine whether *I think* occurs at the end or beginning of a clause, but certain contextual clues were interpreted in favour of one over the other, such as the repetition of the personal pronoun in (74) which indicates that *I think* is part of the frame for the next part of the utterance, and thus counts as a front-positioned occurrence:

(74) *Yeah and I am quite happy I I think sp= specially the phonetics was real fun* (LINDSEI-NO)

The end positioned *I think* were typically easier to determine, like in (75), where *I think* is followed by a conjunction signalling the beginning of a new clause:

(75) *and he's captured her true likeness I think but* (LOCNEC)

In addition to position in the clause, occurrences were also marked for the presence of truncation and markers of hesitation or planning, i.e. discourse markers, filled/unfilled pauses, repetition and truncated words. A great number of occurrences of *I think* are thus represented in both of these categories, since interruptions are often accompanied by hesitation markers:

(76) *erm ... I don't know . I think I think I think yeah <X> .. I probably would like to go back there (LOCNEC)*

(77) *er . I think it's . they al= they always went <'> in the Royal Family <XX> and (LOCNEC)*

5.3.2.1 Results and discussion

Table 5.4 shows the individual proportions of the positions considered in relation to the total number of times *I think* occurs in the extracted samples in total. The overview shows that yet again, word-combinations seem to behave similarly across the native and non-native populations considered in this study. An overwhelming proportion of *I think* in the data selections occur in front position, 81.3 % in LOCNEC, 71.8 % in LINDSEI-SW and 69.1 % in LINDSEI-NO. These findings are compatible with the collocational findings seen in table 5.3, where the majority shows *I think* preceded by a conjunction or a response item, or followed by a word (*that/they/the*) signalling that further content follows.

	LOCNEC		LINDSEI-SW		LINDSEI-NO	
	n	%	n	%	n	%
Front	117	81.3	127	71.8	38	69.1
Mid	3	2.1	10	5.7	4	6.9
End	18	12.5	35	19.8	11	19.0
Relative clause	5	3.5	2	1.1	1	1.7
Front in inserted clause	1	0.7	3	1.7	4	6.9
Σ	**144**	**100**	**177**	**100**	**58**	**100**

Table 5.4: *Position in the clause for I think in LOCNEC, LINDSEI-SW and LINDSEI-NO, raw frequencies and percentages*

However, the higher figure for front position in LOCNEC reflects a slight difference in the distribution of the remaining occurrences. The percentages show that this difference is reflected in *I think* more often occurring in mid- and end position in the learner data than in the NS data. In addition, the relative clause-figure is slightly higher for LOCNEC, and the Norwegian learners show a greater preference for the use of *I think* in both mid-position and in inserted clauses as compared to both the Swedish learners and the native speakers.

Interestingly, the numbers in table 5.4 correspond to a certain extent with Aijmer's (2009) findings on the position of *I don't know/I dunno* in LINDSEI-SW and LOCNEC. Aijmer finds a highly significant difference between the occurrences of *I don't know* in initial position, where the greatest proportion is attributed to the native speakers (Aijmer 2009: 155). In the learner corpus, *I don't know* most often occurs in mid- and end position. Aijmer also finds that *I don't know* very often functions as a speech management signal in LINDSEI-SW, maintaining coherence in the utterance and gaining time for planning ahead (ibid.: 165). In addition, the end position of *I don't know* simultaneously marks the closing of the topic and expresses hedging:

> "When *I don't know* is placed at the end of the turn or utterance it has the function of yielding the floor or fulfilling the desire of the interviewee 'to close a topic' (...) in addition to its attitudinal function to express uncertainty or lack of responsibility" (Aijmer 2009: 155)

It is possible that some of these findings can also be transferred to the functions of *I think*, considering its similar semantic properties as well as its similar frequency distributions.

I think positioned in mid-position or in the front of inserted clauses, as seen in (69-70), might be related to lexical retrieval difficulties or a need to clarify previous

claims, which in turn might be a slightly more prevalent need from a learner perspective. However, the mid-position also seems to fulfil a more general function of marking opinion (78-79), epistemic stance, and belief (80):

(78) *you can . use more variation in English I think <overlap /> than in (...)* (LINDSEI-SW)

(79) *I think some . quite a few work there for free and just . yeah of the teachers and some really good teachers I think that (eh) help them (em) we got to see some classes and . and . listen in th= the but mostly I think they had to . deal with issues with the kids and stuff like that* (LINDSEI-NO)

(80) *(er) I've met my . host parents . I think three or four times since I left them* (LINDSEI-SW)

This function is, however, also found in the occurrences of *I think* in mid-position in LOCNEC:

(81) *it was . it was interesting and the more I think we looked at the adverts we actually learnt something about . the way women . are forced to be* (LOCNEC)

In end-position, which seems to be more prominent in learner speech, *I think*, does seem to function in a similar fashion to *I don't know*:

(82) *<B> (...) they were eating a lot of salad . and fruit fresh fruit every day so it was very nice </B>*

<A> (mhm) </A>

<B> I think </B> (LINDSEI-NO)

(83) *we I (eh) noticed a sign . (eh) it was a big sign . beware pedestrians <begin laughter> could come here you know so it was <end laughter> it was quite strange. I think (LINDSEI-NO)*

(84) *<B> I'm pretty excited about that . I think </B>*

<A> yeah .. <overlap /> about </A> (LINDSEI-SW)

In (82-84), it seems that *I think* functions primarily as a discourse marker, signalling that the speaker wishes to give up the floor or to convey that he/she "has nothing more to say" (Aijmer 2009: 157). This is also signalled by the pauses preceding *I think*, which prompts a reaction from the interviewer. *I think* in these examples refer to personal assessments rather than the stating of facts, but it is still possible that the speakers also intend *I think* to function as a hedge, modifying this assessment: 'this is

just what *I* think'. Such a modifying of content is prominent in the more propositional clauses, where *I think* is tagged on as a disclaimer:

(85) *suddenly they give away this very cheap boat trips so . and yeah it's a very natural place for (em) . for Norwegians to go I think* (LINDSEI-NO)

(86) *(erm) .. <swallows> what else .. I got to travel pretty much while I was there I visited . ten states I think* (LINDSEI-SW)

In the native speaker corpus, *I think* seems to function in a similar way, signalling assessment of the preceding clause and signalling floor-yielding:

(87) *<B> yeah a couple of my friends are into that I think <\B>* (LOCNEC)

(88) *<B> that is the lesson yes you've got to be very careful I think mm <\B>*

<A> mm <\A>

<B> yeah <\B> (LOCNEC)

However, the fact that *I think* occurs less in end position in LOCNEC than in NNS corpora, suggests that this function is not as prominent in native speech.

In initial position, the most frequent position in all three corpora, *I think* occurs mainly as a launcher of utterances, taking part in either the frame or the stem of the utterance, as described in Altenberg (1998):

(89) *(er) I I've never really had that that problem .. myself erm . I don't think I think I get on . well with my parents* (LOCNEC)

In initial position, *I think* also seems to be more often surrounded by hesitation features, which is also found to be the case for *I don't know* (Aijmer 2009: 163). As shown in table 5.5, hesitation features occur around *I think* between 33.3 % and 37.3 % of the time:

	%
LOCNEC	33.3
LINDSEI-SW	37.3
LINDSEI-NO	36.2

Table 5.5: The proportion of *I think preceded or followed by hesitation markers (discourse markers, filled/unfilled pauses, repetition and/or truncated words), measured in percentages*

The figures in table 5.5 are slightly higher for the learner corpora, but it is difficult to say whether much emphasis can be laid on this difference. It is likely that learner

language has a greater proportion of hesitation features in general, which thus will be reflected in a count such as the one presented in this table. However, the figures do show that *I think* often appears in this sort of linguistic environment, which generally strengthens the impression of *I think* as a discourse marker functioning as an utterance launcher and a planner in discourse. However, *I think* still seems to display a stronger connection to its literal meanings than other discourse markers, particularly in the learner corpora, as seen in (90), where *I think* is preceded by filled and unfilled pauses, *I don't know*, and the interrupted clause *it's just,* and seems to take part in the stem of the utterance rather than this more disconnected frame:

(90) *and (eh) .. yeah .. I don't know it's just . I think every Norwegian thinks about Copenhagen and they think summer . cheap food . good food tasty food and (eh) . yeah .. just enjoying <overlap /> life* (LINDSEI-NO)

If *I think* is often part of an interrupted clause, this would go further in suggesting that it is part of a planning process, where the speakers may change their minds half-way through as a result of still being present in this planning process. Table 5.6 shows the proportion of clauses which are interrupted in this data set:

	%
LOCNEC	14.6
LINDSEI-SW	15.8
LINDSEI-NO	8.6

Table 5.6: *The proportion of interrupted clauses with I think, measured in percentages*

It is difficult to explain why the numbers for LOCNEC and LINDSEI-SW differ from the low number in the LINDSEI-NO row. However, this diverging figure might be attributed to the small sample from the corpus. In the cases where the Norwegian learners do change their minds after initiating a clause containing *I think,* the interruption typically happens at the transition between given and new information:

(91) *so I think her stay was . a= she experienced a lot and . about herself and cultures and stuff like that but i= it ended . not very well <laughs>* (LINDSEI-NO)

(92) *but I think she is (eh) not so (eh) she doesn't like the picture . because it looks just like her* (LINDSEI-NO)

In other places, the learner is interrupted by the interviewer, making it difficult to determine whether the clause would have been interrupted had the overlap not taken place:

(93) *<B> <overlap /> because I it doesn't matter for me: where to work actually <overlap /> I think it's (eh) </B>*

<A> <overlap /> and how have you found it being a mature student (eh) well you're not the only one in the class </A> (LINDSEI-NO)

In LINDSEI-SW, a number of the interrupted clauses should perhaps rather have been classified as independent clauses in a frame-like front position:

(94) *but I think I've I've heard at least that you can't stay at home . you can't both work and stay at home* (LINDSEI-SW)

(95) *I think i= I suppose if I would have wanted to paint I would probably have wanted to do it like my father so* (LINDSEI-SW)

This structurally independent *I think* can also be found in LOCNEC:

(96) *and erm I think I mean the plot is just basically about the different kinds of characters you find in* (LOCNEC)

Since some assumptions were made as to the possible formulaic status of *I think it's*, particularly in the learner corpora, the instances of *I think it's* in this small-scale material should perhaps be looked further into in this analysis. As a consequence of being one of the most common collocational patterns with *I think* in LOCNEC, LINDSEI-SW and -NO (4.4, 8.4 and 8.5 occurrences per 10,000 words, respectively), *I think it's* is also a common combination in the limited material considered here. Unsurprisingly, *I think it's* commonly occurs in initial position, and this combination is also very often surrounded by hesitation markers, and appears as part of interrupted clauses:

(97) *so I think it's I think it's the fact that I had a bossy German with me that sort of helped as well* (LOCNEC)

(98) *and: it's: something that I I I think it's more .. well it's further from England* (LINDSEI-NO)

(99) *being . being from Korea adopted . to Sweden I th= I think it's .. this has made my life easier in a way* (LINDSEI-SW)

(100) (erm) . (mm) (eh) <starts laughing> yeah I think <stops laughing> (er) I think I do (er) not (er) . you know consciously or <overlap /> not (er) but (er) . I think it's inevitable to (eh) you know (LINDSEI-SW)

This strengthens the assumption that *I think it's* shows signs of being holistically stored, and a preferred choice to start off an utterance, often one of evaluative nature. This contextual information combined with the fact that the combination is twice as frequent in the learner corpora, provides a basis for suggesting that its formulaic status is more prevalent in learner language.

5.3.3 Qualitative CIA: Summary of Findings

Judging from the collected samples and the categorization of *I think* according to clausal position, there is a preference for employing *I think* in initial position in both native and non-native speech. This thus confirms the impression from the extracted collocational patterns and the quantitative information of *I think* presented in chapter 4. Concerning the status of *I think* as a discourse marker, this also seems to be prevalent in all three corpora, since *I think* is found to be (i.) syntactically detachable from a sentence, (ii.) commonly used in initial position, (iii.) followed by a pause, (iv.) operating at both local and global levels of discourse and (v.) displaying vague meaning. Since there was a greater tendency for *I think* to appear in mid- and end position in the clause in the learner corpora, this might indicate that *I think* is more versatile and independent in learner language than in native language. If this is the case, it would thus support the hypothesis that the greater frequency of *I think* in learner language widens its pragmatic scope and strengthens its position as a holistically stored and versatile formulaic sequence. Since no functions were found to be exclusive to the learner corpora, it is possible that they are inspired by native language use, or that similar pragmaticalization processes are at work also in native speech.

It also seems as if *I think* in learner language often functions without any clear global or discourse marking functions, and rather serve to merely present personal opinions and attitudes to propositions made. This greater tendency for evaluation might, as pointed out by Aijmer (2009) in relation to *I don't know*, be due to "the speaker's unwillingness to commit him- or herself" (Aijmer 2009: 164), but it is also likely from considering the examples above that the learners lack other ways of starting of

utterances or bringing the conversation forward. *I think* seems to be commonly used to evaluate in learner speech, with predicative clauses such as *I think it's a very nice town* being very common. This might be part of a strategy of "stitching together" (Altenberg 1998) both propositional and linguistic content to retain fluency and to actually *have something to say* in the face of being asked to speak in a foreign language.

5.4 Formulaic Sequences: Summary

The findings in 5.3 are thus not conclusive, which might be attributed to the limited size of the data, and to the fact that not every instance was classified according to function(s). It seems that further research is needed to fully uncover the differences and similarities of the functional patterns of *I think* in native and non-native speech. It would also be particularly useful to consider prosodic information for this purpose, as seen in e.g. Aijmer (1997) and Kärkkainen (2003), since intonation and stress serve as essential clues in the endeavour to assign sequences to functions. However, the preceding sections offer some preliminary results, and provide a supplement to the predominantly quantitative data presented in chapter 4. It seems that *I think* is a dynamic and versatile formulaic sequence which is used to perform a number of functions, and that this picture is more striking in learner speech than in the language of native speakers. This, in turn, might be due to a number of features relating to the demands of the learner situation, and to the processes of pragmaticalization by which sequences take on extended meanings and functions.

This chapter has explored further the concept of formulaicity and its relation to the inventory of recurrent sequences in LINDSEI and LOCNEC. Several explanations were suggested regarding the overuse, underuse and diverging functional and structural patterns of some of the recurrent word-combinations found in the NS and NNS corpora. No firm conclusions have been reached, but since corpus research on spoken learner language and corpus studies of formulaic sequences are still fairly recent undertakings, the study has perhaps provided a small contribution to these fields.

6 Concluding Remarks

6.1 Strengths and Limitations of the Approach

The aim of this thesis has been to provide an overview of the formulaic inventory of a variety of native and non-native spoken English, in terms of form, function and contrastive differences and similarities. Since the notion of formulaicity is difficult to pin down, the analysis was conducted so as to attempt to shed light on formulaicity in general, and on how formulaic language occurs and operates within both native and non-native language. One important assumption has been that general cognitive processes (*chunking, categorization, rich memory storage* and *cross-modal associations* (Bybee 2010: 7-8)) are influencing both native and learner language varieties, and that the formal and quantitative differences we find between native and non-native language can primarily be explained in terms of these processes, in addition to various aspects relating to the learner situation.

The results presented in chapter 4 shows some of the particular strengths of the corpus-driven method, since they provide information on a large proportion of data, making it possible to put forward ideas about probabilities which are firmly based on naturally occurring data rather than on small-scale material or intuitive ideas. Within the scope of this thesis, such an extensive overview provides "a good indication of what kinds of expression speakers resort to in on-going discourse" (Altenberg 1989: 136-137), which would not have been possible to present through a manual analysis of a more limited data set. In addition, the method is highly unassuming in its nature, considering recurrent word-combinations which differ widely in structural and semantic properties alongside each other, and the definition of formulaic sequences thus had to be continually discussed and revised according to the data presented. In this way, explanations for quantitative and qualitative differences between the NS and NNS corpora had to be worked out without a backdrop of predetermined categories. Not many controversial choices were made in terms of the word-combinations which were ultimately picked out as subject to further analysis in chapters 4 and 5, but the brief presentation and discussion of some of the word-combinations which are more difficult to classify may be seen to be fruitful in itself, as it challenges our conception

of formulaicity, holistic storage and discourse-pragmatic functions. These frequently occurring word-combinations are perhaps, now that fully compositional word-combinations like *I think* and *I don't know* are increasingly recognized, left alone in the 'peripheral' corners of the phraseological sphere, awaiting future research which might serve to change this picture.

Both the determinants for identification of formulaic sequences and the quantitative methods employed have thus been subject to discussion and revision throughout the analysis. Some of the tentative conclusions made in relation to the quantitative findings were further investigated in the structural and functional analysis of *I think* and its collocations in the last sections of chapter 5. Firmer conclusions on the particular issues addressed here can perhaps ultimately be reached if the quantitative scope is narrowed, or if only certain categories are predominantly considered, such as e.g. markers of vagueness, or epistemic tags. This would also make for an easier comparison with previous studies not necessarily based on corpus methods. At the same time, it can be argued that the vastness of the data and the possibility of spotting general tendencies on the basis of this data, makes for a useful and informative approach, working on its own terms. In combination with more qualitatively based approaches, it seems that the corpus-driven recurrent word-combinations method can provide results which are difficult to obtain through the use of other methods.

Concerning limitations of this particular study, one main issue is the lack of prosodic information as a basis for functional analysis. This might prove to be particularly relevant for the identification of formulaic sequences, and is undoubtedly important in the functional analysis of sequences in context. In addition, more systematic studies should be conducted on the appearance of filled and unfilled pauses in spoken language, in relation to the identification and functions of formulaic sequences. All of these concerns take on a final dimension in the study of learner language, where prosodic features and the presence of pauses would have to be considered on the basis of general properties of non-native spoken language contours. It would also be beneficial for the validity of the results presented if the data was controlled for individual variation, as discussed in section 4.4.1, since formulaic language use can be highly idiosyncratic. In addition, the basis provided for comparison between e.g. the different LINDSEI subcorpora would benefit from a more extensive assessment of

learner proficiency based on internal rather than external criteria, as discussed in section 3.2.2.

6.2 Possible Applications of Findings and Suggestions for Further Research

It was claimed in the introduction to this paper that learner corpus research lies at the crossroads between corpus linguistics, linguistic theory, second language acquisition and foreign language teaching. A study such as the present one may yield results which can fuel theories on learners' processing of recurrent word-combinations in English, and attempts have been made throughout to connect corpus results to existing linguistic theory. In more practical terms, it seems that the corpus-driven approach may also provide information which is useful to learners and teachers of English as a foreign language. Although phraseological approaches should not be "the be-all and end-all of language teaching" (Granger and Meunier 2008: 251), Granger and Meunier believe that "awareness of phraseology in the wide sense should be promoted" in teacher training and among learners (ibid.: 251). In a more narrow sense, it should be useful to present to teachers and learners corpus results which point to particular word-combinations which are likely to be considered as markers of non-nativeness in learner language, as particular points of consideration in the learner process. This is a type of what Granger terms "delayed pedagogical use" (Granger 2009: 20), in which data from learner corpus research is presented "with a view to providing a better description of one specific interlanguage and/or designing tailor-made pedagogical tools which will benefit similar-type learners" (ibid.). Although this endeavour is perhaps more relevant to written language with its distinct rules and norms, it should not be excluded from practice in spoken language, since a larger spoken language repertoire is likely to lead to greater confidence for language learners, which in turn should greatly facilitate both spoken and written language production and communication. In a global sense, it is possible that an early awareness of higher-level units as opposed to an exclusive focus on single words and generative rules, might promote and facilitate the acquisition and use of these units, but this is a question which is still open for further investigation. According to Granger and Meunier, there is a need for "teaching and learning practices which are informed by evidence from descriptive and theoretical linguistic analyses, second

language acquisition research, psycholinguistic findings, and which are validated and assessed in the classroom" (Granger and Meunier 2008: 251).

As mentioned in the beginning of chapter 5, the analysis in chapter 4 left us with perhaps more questions than answers. However, considering the generally co-operative nature of corpus linguistics research, it is possible that some of the quantitative and qualitative results of this study may take on further meaning in combination with other, similar studies. In addition, since limitations of space and scope did not allow for the analysis of many of the interesting functional patterns of recurrent word-combinations in both corpora, these findings may provide useful onsets for further research. Comparison with the other subcorpora of LINDSEI or with comparable corpora of native speaker speech would shed further light on the effects of transfer, and on general tendencies across learner populations, which might further substantiate the claim that the factors which operate in the processing of a foreign language are to a great extent influenced by aspects of the learner situation.

According to Nick Ellis (2008), language patterning and phraseological notions "pervades theoretical, empirical, and applied linguistics" (Ellis 2008: 9). This thesis has sought to understand these connections, through theoretical discussions, the analysis of quantitative data, and the presentation of preliminary results. It seems that theoretically founded phraseological studies making use of corpus data underlines "how language draws on basic cognition, on perception, attention allocation, memory and categorization, that it cannot be separated from these as a distinct, modularized, self-governed entity, that knowledge of language is integrated with our general knowledge of the world, and that language use and language function interact with language structure" (Ellis 2008: 5), and that they thus bring us closer to 'the periphery and the heart of language'.

References

Aarts, Bas 2000. "Corpus Linguistics, Chomsky and Fuzzy Tree Fragments", in Christian Mair and Marianne Hundt (eds.), *Corpus linguistics and linguistic theory: Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999.* Amsterdam & Atlanta: Rodopi, pp. 5-13.

Abrahamsson, Niclas and Kenneth Hyltenstam 2009. "Age of Onset and Nativelikeness in a Second Language: Listener Perception Versus Linguistic Scrutiny", in *Language Learning* 59 (2), pp. 249-306.

Aijmer, Karin 1997. "*I think* – an English modal particle", in T. Swan and O.J. Westvik (eds.), *Modality in Germanic Languages: Historical and comparative perspectives*. Berlin/New York: Mouton de Gruyter, pp. 1-47.

Aijmer, Karin 2001. "*I think* as a marker of discourse style in argumentative student writing", in Karin Aijmer (ed.), *A Wealth of English: Studies in Honour of Göran Kjellmer*. Göteborg: Acta Universitatis Gothoburgensis, pp. 247-258.

Aijmer, Karin 2004. "Pragmatic Markers in Spoken Interlanguage", in *Worlds of words. A tribute to Arne Zettersten. Nordic Journal of English Studies. Special Issue*, 3 (1), pp. 173-190.

Aijmer, Karin 2009. "'So er I just sort I dunno I think it's just because...': A corpus study of *I don't know* and *dunno* in learners' spoken English", in Andreas H. Jucker, Daniel Schreider and Marianne Hundt (eds.) *Corpora: Pragmatics and Discourse*. Amsterdam: Rodopi, pp. 151-168.

Aijmer, Karin 2011. "*Well I'm not sure I think.*.The use of *well* by non-native speakers", in *International Journal of Corpus Linguistics* 16 (2), pp. 232-233.

Altenberg, Bengt 1998. "On the Phraseology of Spoken English. The evidence of Recurrent Word-combinations", in A.P. Cowie (ed.), *Phraseology, Theory, Analysis and Applications*. Oxford: Oxford University Press, pp. 101-122.

Altenberg, Bengt 1989. "Speech as Linear Composition", in Caie, Haastrup, Lykke Jakobsen, Nielsen, Sevaldsen, Specht and Zettersten (eds.), *Proceedings from the Fourth Nordic Conference for English Studies, vol. 1*, Department of English, University of Copenhagen, pp. 133-143).

Altenberg, Bengt and Mats Eeg-Olofsson 1990. "Phraseology in Spoken English: Presentation of a Project", in Jan Aarts and Willem Meijs (eds.), *Theory and Practice in Corpus Linguistics*. Amsterdam: Rodopi, pp. 1-22.

Atkins, Sue and Jeremy Clear 1992. "Corpus Design Criteria", in *Literary and Linguistic Computing* 7 (1), pp. 1–16.

Ball, Catherine 1994. "Automated Text Analysis: Cautionary Tales", in *Literary & Linguistic Computing 9 (4),* pp. 295-302.

Barlow, Michael 1996. "Corpora for Theory and Practice", in *International Journal of Corpus Linguistics* 1, pp. 1-37.

Barlow, Michael 2005. "Computer-based analyses of learner language", in Rod Ellis and Gary Barkhuizen (eds.), *Analysing learner language*. Oxford: Oxford University Press, pp. 335-357.

Baumgarten, Nicole and Juliane House 2010. "*I think* and *I don't know* in English as a lingua franca and native English discourse", in *Journal of Pragmatics* 42, pp.1184-1200.

Biber, Douglas 1986. "Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings", in *Language* 62 (2), pp. 384-414.

Biber, Douglas 1993. "The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation. An Overview of Methodology and Findings", in *Computers and the Humanities* 26, pp. 331-345.

Biber, Douglas and Randi Reppen 1998. "Comparing native and learner perspectives on English grammar: A study of complement clauses", in Sylviane Granger (ed.), *Learner English on Computer.* London: Longman, pp. 145-158.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.

Bloomfield, Leonard 1933/1967. *Language*. London: George Allen & Unwin LTD.

Brand, Christiane and Sandra Götz 2011. "Fluency versus accuracy in advanced spoken learner language: A multi-method approach", in *International Journal of Corpus Linguistics* 16 (2), pp. 255-275.

Bybee, Joan 2009. "Usage-Based Grammar and Second Language Acquisition", in P. Robinson & N. C. Ellis (eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition*. New York and London: Routledge. Kindle Edition.

Bybee, Joan 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Chafe, Wallace 1992. "The Importance of Corpus Linguistics to Understanding the Nature of Language", in Jan Svartvik (ed.), *Trends in Linguistics: Directions in Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991.* Berlin: Mouton de Gruyter.

Chomsky, Noam 1965. *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Chomsky, Noam 1957. *Syntactic Structures*. London: Mouton de Gruyter.

Cowie, A. P. 1998. *Phraseology. Theory, analysis, and applications*. *Oxford Studies in Lexicography and Lexicology*. Oxford: Oxford University Press.

Lexicography and Lexicology. Oxford: Oxford University Press.

Dahlmann, Irina and Svenja Adolphs 2009. "Spoken Corpus Analysis: Multimodal Approaches to Language Description", in Paul Baker (ed.), *Contemporary Corpus Linguistics*. London: Continuum, pp. 125-139.

Dechert, Hans W. 1984. "Second Language Production: Six Hypotheses", in Hans W. Dechert, Dorothea Möhle and Manfred Raupach (eds.), *Second Language Productions*, Tübingen: Gunter Narr Verlag, pp. 211-230.

De Cock, Sylvie 1998. “A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English”, in *International Journal of Corpus Linguistics* 3(1), pp. 59-80.

De Cock, Sylvie 2004. “Preferred sequences of words in NS and NNS speech” in *Belgian Journal of English Language and Literatures (BELL), New Series* 2, pp. 225-246.

De Cock, Sylvie, Sylviane Granger, Goffrey Leech and Tony McEnery 1998. ”An automated approach to the phrasicon of EFL learners”, in Sylviane Granger (ed.), *Learner English on Computer.* London: Longman, pp. 67-79.

Dretske, F. I. 1974. "Explanations in linguistics", in D. Cohen (ed.) *Explaining Linguistic Phenomena*. New York: John Wiley & Sons, pp. 21-41.

Dyvik, Helge 1995. “Språk, språklig kompetanse og lingvistikkens objekt”, in Cathrine Fabricius-Hansen og Arnfinn Muruvik Vonen (red.), *Språklig kompetanse - hva er det, og hvordan kan det beskrives? Oslo-studier i språkvitenskap 11*. Oslo: Novus Forlag, pp. 20-41.

Ellis, Nick C., Rita Simpson-Vlach and Carson Maynard 2008. “Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL” in *TESOL Quarterly* 42 (3), pp. 375-396.

Ellis, Nick C. and Peter Robinson 2009. “An introduction to cognitive linguistics, second language acquisition, and language instruction”, in P. Robinson & N. C. Ellis (eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition*. New York and London: Routledge. Kindle Edition.

Ellis, Nick C. 2008. "Phraseology: The periphery and the heart of language", in Fanny Meunier and Sylviane Granger (eds.) *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins, pp. 1-14.

Ellis, Rod and Gary Barkhuizen 2005. *Analysing learner language*. Oxford: Oxford University Press.

Erman, Britt 1987. *Pragmatic expressions in English: A study of 'you know', 'you see' and 'I mean' in face-to-face conversation.* Stockholm: Almqvist & Wiksell.

Erman, Britt and Beatrice Warren 2000. "The idiom principle and the open choice principle", in *Text* 20 (2), pp. 29-62.

Gernsbacher, Morton Ann and Michael P. Kaschak 2003. "Neuroimaging studies of language production and comprehention", in *Annual Review of Psychology* 54, pp. 91-114.

Gilquin, Gaëtanelle and Sylvie De Cock 2011. "Error and disfluencies in spoken corpora", in *International Journal of Corpus Linguistics* 16 (2), pp. 141-172.

Gilquin, Gaëtanelle, Sylvie De Cock and Sylviane Granger (eds.) 2010. *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM.* Louvain-la-Neuve: Presses universitaires de Louvain.

Granger Sylviane, 1994. "The Learner Corpus: A Revolution in Applied Linguistics", in *English Today* 39, 10, 3, p. 25-29.

Granger, Sylviane 1996. "From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora", in K. Aijmer, B. Altenberg and M. Johansson (eds.), *Languages in Contrast. Text-based cross-linguistic studies*. Lund: Lund University Press, pp. 37-51.

Granger, Sylviane 1998a. "Prefabricated Patterns in advance EFL writing: Collocations and Formulae", in A. Cowie (ed.), *Phraseology, Theory, Analysis and Applications*. Oxford: Oxford University Press, pp. 145-60.

Granger, Sylviane 1998b. "The Computer Learner Corpus: A versatile new source of data for SLA research", in Sylviane Granger (ed.), *Learner English on Computer*. London: Longman, pp. 3-18.

Granger, Sylviane 2008. "Learner Corpora", in Anke Lüdeling and Merja Kytö (eds.), *Corpus Linguistics: An International Handbook.* Berlin and New York: Walter de Gruyter, pp. 259-75.

Granger , Sylviane 2009. “The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation” in Karin Aijmer (ed.), *Corpora and Language Teaching*. Amsterdam: John Benjamins, pp. 121-139.

Granger, Sylviane and Fanny Meunier 2008. "Phraseology in language learning and teaching: Where to from here?", in Fanny Meunier and Sylviane Granger (eds.), *Phraseology in Foreign Language Learning and Teaching*, Amsterdam: John Benjamins, pp. 247-252.

Gries, Stefan Th. 2007. “Exploring variability within and between corpora: some methodological considerations” in *Corpora* 1 (2), pp. 109-151.

Gries, Stefan Th. 2008. “Phraseology and linguistic theory: A Brief Survey”, in Sylviane Granger and Magali Paquot (eds.), *Phraseology: An Interdisciplinary Perspective*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 3-25.

Gries, Stefan Th. 2009. “Corpus-based methods in analyses of Second Language Acquisition Data”, in P. Robinson and N. C. Ellis (eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition*. New York and London: Routledge. Kindle Edition.

Gries, Stefan Th. 2010a. “Corpus linguistics and theoretical linguistics: A love-hate relationship? Not necessarily…” in *International Journal of Corpus Linguistics* 15(3), pp. 327-343. [URL]. Available at: http://www.linguistics.ucsb.edu/faculty/stgries/research/overview-research.html#PublicationsEditing

Gries, Stefan Th. 2010b. “Useful statistics for corpus linguistics” in Aquilino Sánchez and Moisés Almela (eds.), *A mosaic of corpus linguistics: selected approaches*. Frankfurt am Main: Peter Lang, pp. 269-291. [URL]. Available at: http://www.linguistics.ucsb.edu/faculty/stgries/research/overview-research.html#PublicationsEditing

Götz, Sandra and Marco Schilk 2011. “Formulaic sequences in spoken ENL, ESL and EFL: Focus on British English, Indian English and learner English of advanced

German learners", in Joybrato Mukherjee and Marianne Hundt (eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a paradigm gap*. Amsterdam: John Benjamins, pp. 79-100.

Halliday, M. A. K. 2004a. "The Spoken Language Corpus: A Foundation for Grammatical Theory" in K. Aijmer and B. Altenberg (eds.), *Advances in Corpus Linguistics*. Amsterdam: Rodopi, pp. 11-38.

Halliday, M. A. K. 2004b. *An Introduction to Functional Grammar.* 3d ed. London: Arnold.

Hasselgren, Angela 1994. "Lexical teddy bears and advanced learners: a study into the ways Norwegian learners cope with English vocabulary", in *International Journal of Applied Linguistics* 4 (2), pp. 237-260.

Hasselgård, Hilde 2009. "Thematic choice and expressions of stance in English argumentative texts by Norwegian Learners", in Karin Aijmer (ed.), *Corpora and Language Teaching.* Amsterdam: John Benjamins, pp. 121-139.

Herriman, Jennifer and Mia Boström Aronsson 2009. "Themes in Swedish advanced learners' writing in English", in Karin Aijmer (ed.), *Corpora and Language Teaching.* Amsterdam: John Benjamins, pp. 101-120.

Hopper, Paul J. and Elizabeth Closs Traugott 2003. *Grammaticalization* (2nd ed.). Cambridge: Cambridge University Press.

House, Juliane 2009. "Subjectivity in English as Lingua Franca discourse: The case of *you know*", in *Intercultural Pragmatics* 6 (2), pp. 171-193.

Jackendoff, Ray 2002. *Foundations of Language*. Oxford: Oxford University Press.

Johansson, Stig 2001. "Grammar Across Speech and Writing", in W. Vagle & K. Wikberg (eds.), *New Directions in Nordic Text Linguistics and Dicourse Analysis: Methodological Issues*. Oslo: Novus, pp. 45-58.

Johansson, Stig. 2009. “Some thoughts on corpora and second-language acquisition”, in Karin Aijmer (ed.), *Corpora and Language Teaching*. Amsterdam: John Benjamins, pp. 33-44.

Kjellmer, Göran 1991. “A mint of phrases”, in Karin Aijmer and Bengt Altenberg (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, pp. 111-127.

Kjellmer, Göran 2003. “Hesitation. In defence of *er* and *erm*”, in *English Studies* 84 (2), pp. 170-198.

Kärkkainen, Elise 2003. *Epistemic Stance in English Conversation*. Amsterdam: John Benjamins.

Labov, William 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Leech, Geoffrey 2000. “Grammars of spoken English: New Outcomes of Corpus-Oriented Research”, in *Language Learning* 50 (4), pp. 675-724.

Lin, Phoebe M. S. and Svenja Adolphs 2009. “Sound evidence: Phraseological units in spoken corpora”, in A. Barfield and H. Gyllstad (eds.), *Researching collocations in another language: Multiple interpretations*. Basingstoke: Palgrave Macmillan, pp. 34-48.

Lind, Marianne, Inger Moen and Hanne Gram Simonsen 2008. ”Syntactic frames and slot fillers in fluent aphasic speech production: Two Norwegian case studies”, in *NorClinLing 2008 Proceedings from the 1st Nordic Conference of Clinical Linguistics.* Joensuu: University of Joensuu, pp. 71-82.

MacWhinney, Brian 2009. “A Unified Model”, in P. Robinson and N. C. Ellis (eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition*. New York and London: Routledge. Kindle edition.

Meyer, Charles 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Moon, Rosamund 1998. *Fixed Expressions and Idioms in English. A Corpus-Based Approach.* Oxford: Clarendon Press.

Mukherjee, Joybrato 2006. "Corpus linguistics and language pedagogy: the state of the art - and beyond", in Sabine Braun, Kurt Kohn and Joybrato Mukherjee (eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt am Main: Peter Lang, pp. 5-24.

Mukherjee, Joybrato 2009. "The grammar of conversation in advanced spoken learner English", in Karin Aijmer (ed.), *Corpora and Language Teaching.* Amsterdam: John Benjamins, pp. 203-230.

Paquot, Magali, Hilde Hasselgård and Signe Oksefjell Ebeling (forthcoming) "Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora" (ms.).

Pawley, Andrew and F. Syder 1983. "Two puzzles for linguistic theory: nativelike selection and nativelike fluency", in Richards, S.C. & Schmidt, R.W. (eds.), *Language and Communication*. London: Longman, pp. 191-226.

Raupach, Manuel 1984. "Formulae in Second Language Speech Production", in Hans W. Dechert, Dorothea Möhle, Manfred Raupach (eds.), *Second Language Productions*. Tübingen: Gunter Narr Verlag, pp. 114-137.

Read, John and Paul Nation 2004. "Measurement of formulaic sequences", in Norbert Schmitt (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins, pp. 23-35.

Ringbom, Håkan 1998. "Vocabulary frequencies in advanced learner English: a cross-linguistic approach", in Sylviane Granger (ed.), *Learner English on Computer*. London: Longman, pp. 41-52.

Saussure, Ferdinand de 1915/1983. *Course in General Linguistics*. Charles Bally & Albert Sechehaye (red.). London: Gerald Duckworth & Co. Ltd.

Scheibmann, Joanne 2000. "*I dunno*: A Usage-Based Account of the phonological reduction of *don't* in American English Conversation", in *Journal of Pragmatics* 32 (1), pp. 105-124.

Schiffrin, Deborah 1987. *Discourse Markers*. Cambridge University Press.

Schmitt, Norbert and Ronald Carter 2004. "Formulaic sequences in action: An introduction", in Norbert Schmitt (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins, pp. 1-22.

Schmitt, Norbert, Sarah Grandage and Svenja Adolphs 2004. "Are corpus-derived recurrent clusters psycholinguistically valid?", in Norbert Schmitt (ed.), *Formulaic Sequences: Acquisition, Processing and Use.* Amsterdam: John Benjamins, pp. 127-152.

Selinker 1972. "Interlanguage", in *International Review of Applied Linguistics* 10, pp. 209-231.

Sinclair, John 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John 1999a. "A Way with Common Words", in Hilde Hasselgård and Signe Oksefjell (eds.), *Out of Corpora. Studies in Honour of Stig Johansson.* Amsterdam: Rodopi, pp. 157-179.

Sinclair, John 1999b. "The Computer, the Corpus and the theory of Language", in Gabriele Azzaro and Margherita Ulrych (eds.), *Transiti linguistici e culturali. Atti del XVIII Congresso nazionale dell'A.I.A.* Trieste: E.U.T, pp. 1-15.

Smith, Michael Sharwood 1994. *Second language learning: Theoretical Foundations.* London: Longman.

Tottie, Gunnel 2010. "*Uh* and *Um* as sosiolinguistic markers in British English", in *International Journal of Corpus Linguistics* 16 (2), pp. 173-197.

Hopper, Paul J. and Elizabeth Closs Traugott 2003. *Grammaticalization. 2nd ed.* Cambridge: Cambridge University Press.

Wray, Alison 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, Alison 2009. *Formulaic language: Pushing the Boundaries*. Oxford: Oxford University Press.

Corpora used:

LINDSEI-SW

Gilquin, G., S. De Cock and S. Granger (eds.) 2010. *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM.* Louvain-la-Neuve: Presses universitaires de Louvain.

LINDSEI-NO (sample; compilation in progress)

Used by courtesy of Susan Nacey, Hedmark University College.

http://www.uclouvain.be/en-307845.html

LOCNEC

Used by courtesy of Sylvie De Cock, Université catholique de Louvain.

http://www.uclouvain.be/en-cecl-lindsei.html

COCA

Davies, Mark 2008. *The Corpus of Contemporary American English: 425 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/

BNC

The British National Corpus, version 3 (BNC XML Edition) 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available online at http://www.natcorp.ox.ac.uk/

ENPC

Johansson, Stig, Knut Hofland, Jarle Ebeling and Signe Oksefjell Ebeling 1997. *http://www.hf.uio.no/ilos/english/services/omc/enpc*

Appendix

LINDSEI Tasks[22]

LINDSEI

I'd like to interview you informally on things of interest in your life for fifteen minutes. To get the conversation started could you please choose one of the following topics and think about what you are going to say. You should aim to be able to talk for 3-5 minutes. The conversation will then continue informally.

Topic 1: An experience you've had which has taught you an important lesson. You should describe the experience and say what you have learnt from it.

Topic 2: A country you have visited which has impressed you. Describe your visit and say why you found the country particularly impressive.

Topic 3: A film/play you've seen which you thought was particularly good/bad. Describe the film/play and say why you thought it was good/bad.

Please don't take any notes as I would like it to be a spontaneous talk.

[22] Copied text and picture from the task sheets used in the interview sessions for the compilation of LINDSEI-NO at Hedmark University College.

Annex 2: Story for retelling

The four pictures below tell a story. Study the pictures and then make up a story around them.

LINDSEI transcription guidelines[23]

Speaker turns

Speaker turns are displayed in vertical format, i.e. one below the other. Whilst the letter "A" enclosed between angle brackets always signifies the interviewer's turn, the letter "B" between angle brackets indicates the interviewee's (learner's) turn. The end of each turn is indicated by either </A> or </B>.

Overlapping speech

The tag <overlap /> (with a space between "overlap" and the slash) is used to indicate the beginning of overlapping speech. It should be indicated in both turns. The end of overlapping speech is not indicated.

Punctuation

No punctuation marks are used to indicate sentence or clause boundaries.

Empty pauses

Empty pauses are defined as a blank on the tape, i.e. no sound, or when someone is just breathing. The following three-tier system is used: one dot for a "short" pause (< 1 second), two dots for a "medium" pause (1-3 seconds) and three dots for "long" pauses (> 3 seconds).

Filled pauses and backchannelling

Filled pauses and backchannelling are marked as (eh) [brief], (er), (em), (erm), (mm), (uhu) and (mhm). No other fillers should be used.

Unclear passages

A three-tier system is used to indicate the length of unclear passages: <X> represents an unclear syllable or sound up to one word, <XX> represents two unclear words, and <XXX> represents more than two words.

If transcribers are not entirely sure of a word or word ending, they should indicate this by having the word directly followed by the symbol <?>.

Unclear names of towns or titles of films for example may be indicated as <name of city> or <title of film>.

Anonymisation

Data should be anonymised (names of famous people like singers or actors can be kept). Transcribers can use tags like <first name of interviewee>, <first name and full name of interviewer> or <name of professor> to replace names.

Truncated words

Truncated words are immediately followed by an equals sign.

Spelling and capitalisation

British spelling conventions should be followed. Capital letters are only kept when required by spelling conventions on certain specific words (proper names, I, Mrs, etc) – not at the beginning of turns.

Contracted forms

All standard contracted forms are retained as they are typical features of speech.

Non-standard forms

Non-standard forms that appear in the dictionary are transcribed orthographically in their dictionary accepted way: *cos, dunno, gonna, gotta, kinda, wanna* and *yeah*.

Acronyms

---

[23] Copied from University of Louvain, *Centre for English Corpus Linguistics* [URL], http://www.uclouvain.be/en-307849.html

If acronyms are pronounced as sequences of letters, they are transcribed as a series of upper-case letters separated by spaces.

If, on the other hand, acronyms are pronounced as words, they are transcribed as a series of upper-case letters not separated by spaces.

Dates and numbers

Figures have to be written out in words. This avoids the ambiguity of, for example, "1901", which could be spoken in a number of different ways.

Foreign words and pronunciation

Foreign words are indicated by <foreign> (before the word) and </foreign> (after the word).

As a rule, foreign pronunciation is not noted, except in the case where the foreign word and the English word are identical. If in this case the word is pronounced as a foreign word, this is also marked using the <foreign> tag.

Phonetic features

(a) Syllable lengthening

A colon is added at the end of a word to indicate that the last syllable is lengthened. It is typically used with small words like *to*, *so* or *or*. Colons should not be inserted within words.

(b) Articles

-when pronounced as [ei], the article *a* is transcribed as a[ei];

-when pronounced as [i:], the article *the* is transcribed as the[i:].

Prosodic information: voice quality

If a particular stretch of text is said laughing or whispering for instance, this is marked by inserting <starts laughing> or <starts whispering> immediately before the specific stretch of speech and <stops laughing> or <stops whispering> at the end of it.

Nonverbal vocal sounds

Nonverbal vocal sounds are enclosed between angle brackets.

Contextual comments

Non-linguistic events are indicated between angle brackets only if they are deemed relevant to the interaction (if one of the participants reacts to it, for example).